University of Alberta

Development of an Obstetrical Outcome Measure to Assess Morbidity in Newborns - Newborn Morbidity Index

by

Anila Verma ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Medical Sciences-Medicine

Edmonton, Alberta

Fall 1997

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-23084-8

Canada

University of Alberta

Library Release Form

Name of Author: Anila Verma

Title of Thesis: Development of an Obstetrical Outcome Measure to Assess Morbidity in Newborns - Newborn Morbidity Index

Degree: Doctor of Philosophy

Year this Degree Granted: 1997

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

September 26, 1997

1718 Arens Road
Gardiner Park, Regina
Saskatchewan S4V 1H4

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Development of an Obstetrical Outcome Measure to Assess Morbidity in Newborns - Newborn Morbidity Index submitted by Anila Verma in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Medical Sciences - Medicine.

R. Sauve, FRCP(C)

B.F. Mitchell, FRCS(C)

M. Harrison, Ph.D.

T.O. Maguire, Ph.D.

G.Baker, ~~FRCS(C)~~ Ph. D

N.B. Okun, FRCS(C)

D.M. Olson, Ph.D.

September 25, 1997

To my husband, Mukesh,
for his undaunting patience, encouragement, love and support

and

To my kids, Kshitij and Rajat,
for their extreme understanding and courage

Abstract

The purpose of this research project was to investigate, develop and recommend a scaling model and a scoring system for a discriminative obstetrical outcome measure of morbidity in newborns - Newborn Morbidity Index. The proposed tool will serve to compare alternate obstetrical therapeutic strategies and assess the impact of maternal diseases on adverse events in newborns. The development of the tool involved substantive, structural and external validation, and recommendation of a scoring system for its field application. Substantive validation resulted in a proposal of a morbidity index consisting of twenty-two attributes of morbidity subdivided by their levels of severity into sixty-six binary items. Four hundred and eleven newborns at $\geq$28 weeks gestation with complete maternal and newborn records were recruited from the intensive care and regular nurseries at the Royal Alexandra Hospital in Edmonton. Data on newborn morbidity items and maternal features were collected. Structural validation and scale modeling were performed within the framework of Item Response Theory. Dimensionality testing established that there was a single dimension underlying the data set. The goodness-of-fit criteria confirmed that the data set conformed to latent trait theory models. Both one and two-parameter models of the Item Response Theory were employed for structural validation and scale modeling. The sequence and scaling of morbidity items using the one-parameter model were found to be more agreeable with regard to the clinical judgment. Therefore, the scale values from the one-parameter model were recommended for external validation. After exploring three different scoring systems, recommendations to use an aggregated index value of the highest five items scored by a newborn on the morbidity scale were made. The criteria used for cross-

sectional and longitudinal external validation further supported the validity and usability of the proposed 'Five-Item' scoring system for the Newborn Morbidity Index. The sequencing and scaling of items can be refined in future studies by employing larger sample sizes to ensure adequate number of newborns in each category of a morbidity item. This plan will furnish more stable, robust and clinically sound parameter estimations. Future field-testing will provide evidence of desired performance of the tool in both clinical and population studies.

Acknowledgment

I wish to express sincere appreciation and gratitude to the members of my academic advisory committee for their thoughtful contributions to this work. Dr. B.F. Mitchell, my mentor, advisor and guide extended leadership and clinical expertise to the project. I will remain indebted to him for inspiration, focused support and insight. Dr. T.O. Maguire introduced me to the exciting field of measurement and psychometrics, and coached me in the methodological and analytic techniques involved. I am grateful to him for his consistent guidance and inspiration. I wish to acknowledge that both Dr. Mitchell and Dr. Maguire were responsible for the growth and evolution of the research project.

A special thanks to Dr. D. M. Olson, the co-supervisor, for his continued guidance towards problem solving and planning during the course of the study. I wish to express deep appreciation for his assistance in helping me to preserve my perspective. I thank Dr. N. Okun for her supervisory role, constructive criticism and advice. I wish to take this opportunity to express my gratitude to Dr. W. A. McBlain for supporting and accenting the importance of integrity in research.

I am thankful to the Neonatologists, especially Drs. P. Etches and A. Peliowski for assisting in the preparation of the item-list for the proposed index. I also wish to thank all the Neonatologists and Pediatricians with admitting privileges at the Royal Alexandra Hospital for their assistance with the concurrent validation aspect of the study.

I also wish to express sincere thanks to the following individuals: Wendy Tomuschat, for data collection and data entry; Darlene Cribb and Florence Miskew deserve special acknowledgment for being efficient and considerate support team; the staff at the Department of Health Records, Royal Alexandra Hospital for accommodating this study in their busy schedule; the investigators of the SNAP and the NTISS study at Vancouver, especially the study coordinator, Joanna Sales, for forwarding data on these scores for the Royal Alexandra Hospital site.

Finally, I wish to thank the Perinatal Research Centre, University of Alberta for its assistance with studentship during the initial phase of this research. The Hospital for Sick Children Foundation and the MSI Foundation are acknowledged for their support of the research project.

TABLE OF CONTENTS

	Page
CHAPTER 1	
INTRODUCTION	
1.1 Rationale	1
1.2 Purpose	3
1.3 Specific Objectives	3
1.4 Aims of Newborn Morbidity Index	3
1.5 Rationale of Validation of Newborn Morbidity Index	4
1.6 Rationale of Scaling of Newborn Morbidity Index	5
1.7 Underlying Assumptions	5
CHAPTER 2	
REVIEW OF SELECTED LITERATURE	
2.1 Review of Newborn Morbidity	6
2.1.1 Definition	6
2.1.2 Prevalence of Morbidity at Birth	7
2.1.3 Newborn Morbidity, Birth Weight and Gestational Age	8
2.2 Review of Available Tools in Maternal-Child Health	9
2.3 Review of Issues in the Development and Construction of Health Status Outcome Measures	15
2.3.1 Approaches to Development of Measures of Morbidity	15
2.3.2 Classification of Health Status Measures by their Purpose	17
2.3.3 Implications for Development of Purpose-Specific Health Status Measures	18

2.4 Review of the Models of Structural Validation and Scaling..... 20
2.4.1 Classical Test Theory versus Item Response Theory.... 20
2.4.2 Item Response Theory Model..... 22
2.4.2.1 Configuration of the Model for Dimensionality 22
2.4.2.2 Item Analysis and Parameter Estimation..... 25
2.4.2.3 Evaluation of the Degree of Fit of Items..... 34
2.4.2.4 Recommendation of Scaling and Scoring Models 35

CHAPTER 3

SUBJECTS AND METHODS

3.1 Substantive Validation..... 37
3.1.1 Analysis of Literature..... 37
3.1.2 Item Generation and Item Review..... 41
3.1.3 Preliminary Item Testing..... 45

3.2 Structural Validation and Scaling..... 49
3.2.1 Study Site and Sample Size..... 50
3.2.2 Study Design and Study Sample..... 50
3.2.3 Inclusion Criteria..... 51
3.2.4 Data Collection Strategies..... 52
3.2.5 Recruitment of Subjects..... 54
3.2.6 Data Collection..... 55

3.3 External Validation..... 56
3.3.1 Events for Cross-Sectional External Validation..... 56
3.3.2 Events for Longitudinal External Validation..... 59

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1. Descriptive Statistics..... 62

4.1.1 Subjects..... 62

4.1.2 Summary of Modifications to the Item-List..... 64

4.1.3 Descriptive Statistics: Items..... 69

4.2 Inferential Statistics..... 74

4.2.1 Dimensionality of the Data Set..... 74

4.2.2 Structural Validation and Scaling of the Items..... 75

4.2.2.1 Item Discrimination Parameter..... 82

4.2.2.2 Item Difficulty Parameter and Scaling..... 86

4.2.2.3 Estimate of Morbidity (θ)..... 89

4.2.2.4 Evaluation of the Degree of Fit..... 97

4.2.2.5 Scoring System..... 107

4.2.3 External Validation..... 109

4.2.3.1 External Validation: Continuous Measures..... 109

4.2.3.2 Distribution of the Five-Item Score..... 112

4.2.3.3 External Validation: Categorical Measures..... 113

4.2.3.4 Representative Population and Maternal Events 114

CHAPTER 5

SUMMATION, UTILIZATION, STRENGTHS, LIMITATIONS

AND FUTURE IMPLICATIONS

5.1 Summation..... 117

5.2 Utilization..... 121

5.2.1 Properties of NMI..... 121

5.2.2 Salient Features..... 122

5.2.3 Performance Characteristics..... 123

5.2.4 Future Plan for Utilization..... 123

5.3 Strengths and Limitations 124
5.3.1 Strengths 124
5.3.2 Limitations 124

5.4 Future Implications 125
5.4.1 Pertaining to Design 125
5.4.2 Pertaining to Items 125
5.4.3 Pertaining to External Validation 126
5.4.4 Pertaining to Psychometrics and Applied Measurement 126

REFERENCES 127

APPENDICES 136

LIST OF APPENDICES

Appendix	Page
1. Definitions of Measurement Terms	136
2. Assessment of Items by Physicians	137
3. Visual Analogue Scale – Concurrent Validation Form	139
4. Notice of Information to Physicians	140
5. Form of Ethics Approval	142
6. Guidelines for Data Collection and Operational Definitions	143
7. Newborn Data Collection Form	146
8. Maternal Data Collection Form	149
9. Case Contact Client Data Form	150
10. Item Function Curves for 50 Items in the Index	152
11. Item Function Curves for Items with Poor-Fit	177
12. Form for Data Collection: 50-Item Newborn Morbidity Index	181

LIST OF TABLES

Table		Page
2-1.	Issues in the Construction of Health Status Measures.....	19
3-1.	Specifications of Manifest Items of Newborn Morbidity.....	39
3-2.	Morbidity Items in Ascending Order by Average Weight.....	43
3-3.	Provisional Morbidity Item-List	46
4-1.	Newborns: Birth Weight and Gestational Age.....	63
4-2.	Newborns: Percentile Values.....	63
4-3.	Maternal Characteristics.....	63
4-4.	Frequency Distribution of Maternal and Delivery Characteristics	64
4-5.	Summary of Modifications to the Item-List.....	65
4-6.	Raw Score Frequency of Items.....	70
4-7.	Frequency Distribution of the Number of Items Scored.....	73
4-8.	DIMTEST Statistics for Determination of Dimensionality.....	74
4-9.	Goodness-of-fit Criteria.....	77
4-10.	Performance Characteristics of Models.....	81
4-11.	Discrimination Parameter: Two Parameter Model.....	85
4-12.	Morbidity Items in Ascending Order by Difficulty Parameter....	87
4-13.	Frequency Distribution, Percentile Values and 95% CI: NMI.....	91
4-14.	Morbidity Items in Ascending Sequence by Scale Values One Parameter Model.....	93
4-15.	Morbidity Items in Ascending Sequence by Scale Values Two Parameter Model.....	95

4-16. Difficulty Parameter - Three Groups by Gestational Age..... 100

4-17. Pearson's Correlation Coefficient: Sets by Gestational Age..... 102

4-18. Correlation: NMI and Scoring Systems..... 108

4-19. External Validation Events: Continuous Variables..... 110

4-20. Descriptive Statistics: Five-Item Score..... 112

4-21. External Validation Events: Categorical Variables 113

4-22. Five-Item Score and Maternal Disease Conditions..... 115

LIST OF FIGURES

Figure		Page
2-1.	Item Characteristic Curves.....	28
4-1.	Frequency Distribution of Items Scored.....	73
4-2.	One Parameter Model of IRT: BILOG.....	80
4-3.	Two Parameter Model of IRT: BILOG.....	80
4-4.	Frequency Distribution of Morbidity: One-Parameter Model..	90
4-5.	Frequency Distribution of Morbidity: Two-Parameter Model..	90
4-6.	Content versus Test Based Morbidity Estimates One Parameter Model.....	98
4-7.	Content versus Test Based Morbidity Estimates Two Parameter Model.....	98
4-8.	Item Difficulty Parameter: Scatter Matrix Plot of 3 Sets.....	103
4-9.	Morbidity Estimates Group A: Set 1 vs. Set 2.....	104
4-10.	Morbidity Estimates Group B: Set 1 vs. Set 2.....	105

LIST OF SYMBOLS AND ABBREVIATIONS

Binary Item Number	BIN
Chi-Square	Chi-Sq.
Confidence Interval	CI
Difficulty parameter for item i	b_i
Discrimination Parameter for Item i	a_i
Item Characteristic Curve	ICC
Item Response Theory	IRT
Length of Stay	LOS
Morbidity	θ
Newborn Morbidity Index	NMI
Neonatal Intensive Care Unit	NICU
Neonatal Therapeutic Intensity Sub Score	NTISS
Pearson's Correlation Coefficient	r
Public Health Nurse	PHN
Score for Neonatal Acute Physiology	SNAP
Serial Number	S.No.
Small for Gestational Age	SGA
Standard Deviation	SD
Standard Error of Estimate	SEE

Chapter 1

Introduction

1. Rationale for the Study

In developed countries, advances in maternal-child health have contributed to a decline in perinatal mortality [1,2]. Approximately 15-20% of all newborns have some degree of morbidity at birth, ranging from minimal to severe. For very premature newborns (e.g. 22-26 weeks gestation), improving survival rate is an important goal of care, whereas the emphasis of care for newborns born around 28 weeks of gestation shifts towards reduction of morbidity at birth. Accordingly, in near-term pregnancy, significant morbidity at birth rather than mortality is a more appropriate outcome measure for assessing perinatal interventions or maternal disease conditions.

Until recently, low birth weight, low gestational age and low Apgar score were often considered to be sufficient proxy outcome measures reflecting newborn morbidity after adverse obstetrical events [3-7]. However, birth weight and gestational age are intermediate outcomes, and are too crude to detect subtle differences in newborn morbidity that most obstetrical interventions or therapies might produce. The Apgar score [8-11] frequently used as a measure of outcome of obstetrical practice, was originally designed as a discriminative index to identify newborns in need of immediate cardiopulmonary resuscitation. Though it performs well as an evaluative measure of an infant's condition in the first ten minutes after birth, the score is predominantly a

reflection of acute intrapartum events and provides insufficient information about the newborn's health status beyond the neonatal period [9,10,12,13]. In the absence of appropriate global measures of morbidity, researchers often use admission to the neonatal intensive care unit (NICU) or length of stay in the NICU as surrogate outcome measures following perinatal interventions or maternal disease conditions, in spite of acknowledging their limited roles as true reflections of newborn morbidity [14,15].

Specific and individual newborn morbidity attributes, such as apnea, oral feeding difficulty, hypoglycemia etc. continue to be important for obstetricians to address specific research questions. However, these individual attributes fail to reflect the overall neonatal health/morbidity status at birth. In addition, since most interventions are expected to produce only small differences in the major individual neonatal morbid outcomes, large sample sizes are required to detect clinically important differences. Currently, there is no available standardized, validated tool for assessing general newborn morbidity, which can serve to reliably discriminate between the two arms of obstetrical interventions or therapeutic strategies. Without a standardized obstetric discriminative outcome measure, treatment programs aimed at reducing newborn morbidity remain difficult to evaluate.

To be widely useful, such a tool should be based on pathophysiological attributes of clinical examination and be easily administered. Additionally, the included individual attributes of newborn morbidity should be relevant to most maternal conditions and settings. We anticipate that the proposed outcome measure will not only provide a discriminative assessment of treatments due to divergent perinatal practice patterns but

also be a useful tool in perinatal epidemiology to estimate the burden of illness due to adverse maternal factors on newborn morbidity.

2. Purpose of the Research Project

The purpose of this research project was to investigate, develop and recommend a scaling model and scoring system for a discriminative obstetrical outcome measure. The proposed outcome measure is designed to allow estimation and comparison of the effectiveness of obstetrical interventions and therapeutic strategies on morbidity in newborns at greater than 28 weeks gestational age.

3. Specific Objectives of the Study

The specific objectives of the project were to develop and validate an obstetrical outcome measure, named Newborn Morbidity Index (NMI), and to recommend a scaling model and scoring system for the NMI.

4. Aims of Newborn Morbidity Index

The primary aim of the NMI is to serve as a fine discriminative index of newborn morbidity following obstetrical interventions, particularly in the range of mild to moderate morbidity. In perinatal epidemiology, the NMI can serve to provide assessment of the impact of maternal disease conditions or adverse events on newborn morbidity.

5. Rationale for Validation of Newborn Morbidity Index

Prior to widespread implementation of a tool or outcome measure, an evaluation of the *relevance*, *representativeness* and *responsiveness* of the items selected, and establishment of their validity and reliability are essential [16]. Issues such as the number of dimensions (or constructs) underlying the data, the sequencing of items, and the scaled distances between items are addressed during the development of an outcome measure, to provide for an accurate representation of the latent trait (of newborn morbidity) as it really exists. Often the selected attributes or items have been designed to capture solely the trait(s) of interest (e.g. morbidity in our proposed index), with little or no contamination by other dimensions or constructs.

6. Rationale for Scaling Newborn Morbidity Index

A model of scaling has a bridging function between the data, on the one hand, and the substantive theory, on the other. Van der Ven [17] defines a scaling model as a numerical relational system, and measurement as the representation of an empirical relational system in or by a numerical relational system. Scaling may be defined as an attempt to find a set of coherent rules that map observable characteristics onto a numerical scale in a manner consistent with the theory, so that the representation by the scores is meaningful statistically and otherwise [18].

The integrative nature of the attributes of morbidity in newborns coupled with the level-

of attribute approach to explaining morbidity provides an opportunity to develop a quantitative outcome measure describing the degree of newborn morbidity at birth.

7. Underlying Assumptions

Morbidity in newborns is integrative and global in nature. The proposal of developing an index of newborn morbidity is based on the underlying assumptions that the trait of newborn morbidity is unidimensional and falls along a continuum, from none to high.

Chapter 2

Review of Selected Literature

The literature review has been organized into four sections. The first is concerned with theoretical views of newborn morbidity, the second is a review of the available tools in maternal-child health, the third discusses issues in the development and construction of health status outcome measures, and the fourth is a review of the models of structural validation and scaling.

2.1 Review of Newborn Morbidity

2.1.1 Definition

Birth is an obligatory change of environments for the fetus. At transition from the intrauterine to extrauterine existence, each organ-system of the fetus must adapt to the changed environment. For a very small percentage of infants, transition is never achieved; for a slightly larger number, transition is delayed or complicated, but for most, transition is so smooth as to appear uneventful. Complicated transition is what has been called newborn morbidity by Desmond et al. [13]. Implicit in the work of Desmond is the notion that morbidity itself is a single continuum running from nil to high [19]. J.S. Drage and H. Berendes, Virginia Apgar, [20,21,23] and Molly E. Towell [22] in the 1960s explained that neonatal mortality and morbidity are each a manifestation in sequence, resulting from some underlying pathology in the newborn. They stated that

newborn morbidity, instead of mortality is a more informative criterion of the potential hazards or effects due to maternal and fetal diseases, and intrapartum influences on the newborn's health status [24].

Newborn morbidity (a loss of normal somatic/brain growth and/or function) is reflected in the attributes of pathophysiology at birth. As postulated by Knaus et al. [25], regardless of disease etiology, the more disturbed a patient's physiology, the sicker or more morbid the patient. Adverse events or morbidity in newborns at birth are attributed to one or more of the three causes: maternal disease states, fetal disease states, and antepartum/intrapartum events.

2.1.2 Prevalence of Morbidity at Birth

Approximately 15-20% of all newborns have some degree of morbidity at birth, ranging from minimal to severe. Of these morbid newborns, 85-90% are admitted to the neonatal intensive care unit (NICU) and the rest may be managed with regular newborn care. The incidence of moderate to severe morbidity in term infants is 4% [26]. Intrapartum accidents and congenital anomalies are the major causes of morbidity in term babies. Other contributors are infections, asphyxia, meconium aspiration, and isoimmunization. TW Kurczynski [27] reported that approximately 2% of the newborns might have a serious congenital anomaly necessitating immediate intervention. Minor congenital anomalies make a negligible contribution to morbidity at birth. There is a natural overlap between minor anomalies and normal variants, and the latter are arbitrarily distinguished

by prevalence greater than 4%. The overall incidence of in-born metabolic disease is estimated to be about 1 in 4000 newborns [28]. In certain rare conditions of in-born errors of metabolism, or congenital anomalies involving internal organs, morbidity may be occult at birth, and becomes apparent only in the post-neonatal period. Morbidity due to such conditions may not be accounted for in the spectrum of morbidity at birth, and hence will not be reflected in our proposed outcome measure. We will score morbidity from birth until 7 days after birth or discharge from the hospital (whichever is earlier).

2.1.3 Newborn Morbidity, Birth Weight and Gestational Age

Generally, for any given gestational age, the lower the birth weight, the higher the morbidity; and for any given birth weight, the shorter the gestational age, the higher the neonatal mortality/morbidity. The lowest risk of neonatal mortality occurs among infants with birth weights of 3000-4000 g and gestational ages of 38-42 weeks [26]. With each increasing week of gestational age, the neonatal survival rate increases and neonatal morbidity decreases as also illustrated in the following table:

(From the Technical Bulletin No. 133, American College of Obstetrics and Gynecology, 1989)

Gestational Age (weeks)	Approximate Mean birth weight (g)	Chance of survival (%)	Chance of Survival free of major morbidity (%)
22	500	0	0
23	575	4	2
24	670	17	9
25	775	30	18
26	900	51	41
27	1,025	64	54
28	1,150	75	67
29	1,250	81	74
30	1,400	87	81
31	1,550	93	87
32	1,750	95	90
33	2,000	97	93
34	2,200	98	95
35	2,400	99	97
36	2,600	99+	98

2.2 Review of Available Tools in Maternal-Child Health

In the last four decades, approximately 30 scales and inventories have been devised for assessing newborns. Most of the indices were developed to predict developmental outcome or mortality in the low-birth weight and/or pre-term newborns [29-32]. Only a few of the available predictive outcome measures have been thoroughly validated, and even fewer have included a sample size of more than 100 subjects for validation. The existing inventories can be broadly classified into four categories: Obstetric Risk Scores, Predictors of Developmental Outcome, Predictors of Acute Mortality Risk, and Morbidity and Mortality Indexes.

Obstetric Risk Scores: Based on the antepartum and intrapartum events, and newborn characteristics, these scores were first developed in the 1970s to recognize high-risk pregnancies requiring closer monitoring [33-35]. Their principal purpose was to act as screening tools for antenatal and neonatal referral and not as measures of newborn morbidity. These scores including the recently developed Manitoba Obstetric Scoring System serve to classify the obstetric population into broad categories by their risk factors, and do not provide any discrimination in outcomes following obstetrical interventions.

Predictors of Developmental Outcome: The Sarnat Score [36], Bayley's Developmental Score [37] and Neuro-Biological Risk Score (NBRS) [38] are examples of measures in this category. Although the staging system of Sarnat and Sarnat was

developed on an extremely limited sample (n=21) for a single disease state, it is widely used as a clinical and electro-encephalographic descriptor of neurological prognosis following newborn ischemic encephalopathy. Some of the development outcome measures are extremely good predictors of long-term sequelae. Most often these measures are based on items of therapies and diagnoses at discharge from the hospital, and are focused on severe morbidity or disorders involving the nervous system. The study of behavioural states in infants, such as the NBRS has attracted wide interest as an indicator of the functional integrity of the central nervous system during the fetal, neonatal, and infant periods of development. These scales essentially prognosticate integrity of the nervous system and attempt to predict long-term developmental outcomes. Due to their relatively narrow focus, only on the neurological parameters, these measures are not appropriate for assessing comprehensive and global newborn outcomes for obstetrical studies.

Predictors of Acute Mortality Risk: The Neonatal Therapeutic Intensity Scoring System (NTISS) [40,41], the Score for Neonatal Acute Physiology (SNAP) [42,43] and the Clinical Risk Index for Babies (CRIB) [43] are the most recent examples of these scores. The NTISS is a score of therapeutic intensity in the NICU, and is a reflection of the quality of care and resource utilization. The SNAP developed as a predictor of neonatal mortality is an organ system physiology-based illness severity index designed for the neonatal intensive care population. It is based on the same explicit scoring system as the Physiological Stability Index (PSI) [44] and consists of 26 objective physiological and biochemical measurements derived from 34 tests and vital signs. It is scored at 24

hours of age. The SNAP correlates with neonatal mortality in the NICU populations, corresponds closely to physician and nursing estimates of mortality risk and predicts hospital costs and length of stay [45]. The CRIB is a simpler, 6-item physiology-based severity score, and was developed solely for infants <31 weeks' gestation or <1,500 g, or both. The SNAP is laboratory-intensive and due to its composition is likely to result in low discrimination among newborns having less than severe morbidity. In contrast, the aim of our proposed measure is to be a sensitive and reliable discriminative index in the range of mild and moderate morbidity in newborns at >28 weeks gestation.

In our proposed outcome measure, NMI, we have not included most items of biochemical or physiological testing, since their inclusion could focus the index on dimensions other than morbidity, making it dependent on the context of measurement of those diagnostic tests. The rationale for excluding them from our item list include the following considerations:

a) False-positive or false-negative results of biochemical tests could result in spurious correlation among the items and hence confound conclusions.

b) Variations in the availability of sophisticated tests among NICUs and variations in the management strategies among treating physicians can introduce bias in measurement.

c) The outcome measure will become laboratory intensive, increasing the expense and (possibly) decreasing use.

d) In obstetrical interventions, most of the newborns of interest may not be required to undergo a number of tests and evaluations. Therefore, the inclusion of those items of laboratory testing will preclude the usage of this tool in many obstetrical studies.

Perinatal Morbidity and Mortality Index: An inventory of newborn morbidity items and mortality was developed in 1985 by Hannah et al. to serve as a discriminative measure of perinatal outcome, and has been used in at least two clinical trials [46,47]. Hannah et al. were the first to propose and apply an index of newborn morbidity and mortality as an obstetrical outcome measure, constituted predominantly of items of adverse pathophysiological state. However, the psychometric properties, and the scaling criteria of that index were not ascertained. We know of no other studies that have attempted to further refine and validate the index as proposed by Hannah et al.

Clearly, a large number of outcome measures or tools have been developed to assess various aspects of maternal child health. However, none of the available inventories or scales matches the characteristics or specifications of our proposed discriminative obstetrical outcome measure designed psychometrically. These characteristics include:

(a) Predominantly pathophysiology-based: Being predominantly pathophysiology based, the proposed outcome measure would be less subject to variations over changing times, therapeutic strategies and settings as opposed to a measure that is a mixture of the attributes of diagnosis, risk-factors and therapy.

(b) Simple to use: A limited number of routinely performed tests and clinical assessments of newborns are included. Therefore, the measure would be easily administered, less time-consuming and not dependent on rare, expensive or invasive testing.

(c) Valid: Validation will include substantive, structural and external validation of the proposed index. The validation procedure requires empirical investigations, with the nature of the tool and the form of validity dictating the needed form of evidence. It is important to emphasize the iterative process of validation; repeated demonstrations of expected associations between the score on the morbidity index and functions or aspects external to the trait enhance the utility of the outcome measure.

(d) Explicit: There would be an explicit reporting on performance characteristics of the NMI including measures of internal consistency and standard error of measurement.

(e) Generalizable: The proposal is to employ Item Response Theory (IRT) for item calibration and scale modeling. The property of invariance of item and morbidity parameters is the cornerstone of the IRT. This property implies that the parameters that characterize an item do not depend on the morbidity distribution of the newborn population, and the parameter that characterizes a newborn does not depend on the set of test items provided in the index. Therefore, the proposed outcome measure would be generalizable to different obstetrical settings, situations and conditions for newborns at greater than 28 weeks gestation. Newborns at less than 28 weeks gestation were excluded because of the following three considerations: First, there are major fetal developmental changes around 28 weeks gestation leading to substantial differences in pathophysiological parameters at birth; Second, researchers in the past have proposed and validated outcome measures relevant for newborns at less than 28 weeks gestation, such as, Morbidity Index by Minde et al [30], CRIB [43] etc.; Third, more than 95% of the

newborn population is born at greater than 28 weeks gestation, and a number of current obstetrical trials involve women delivering at greater than 28 weeks gestation. For such trials, there are no standardized outcome measures of newborn morbidity consisting of clinically relevant items of pathophysiology.

2.3 Review of Issues in the Development and Construction of Health Status Outcome Measures

2.3.1 Approaches to Development of Measures of Morbidity

There are four general approaches that are used to measure illness severity or morbidity. The approaches are:

Diagnosis based: Some of the adult illness measures are based on diagnosis and co-morbidities [48], and disease staging [49], e.g. the Federal Diagnosis-Related Groups (DRGs) is a classification of illness severity (as measured by hospital costs) by diagnosis and diagnostic procedures [50]. A scoring system like the DRG is clinically plausible, readily available and applicable to all hospital admissions. The main limitations are variations in the severity of disease for the same diagnosis and inter-hospital and inter-clinician variations in the diagnostic classification. In addition, the diagnosis itself could be an indicator of the efficiency and accuracy of an ever-evolving medical system in addition to being an indicator of morbidity.

Risk-Factor based: A number of measures are based on symptoms, physical findings, events or demographic characteristics which are markers of risk. Examples include trauma scales [51-52], burn scales [53], myocardial infarction indices [54], antenatal and intrapartum scoring indices [33-35], and the Mortality Prediction Model [55]. Their focus on specific disease states or organ systems makes them applicable in specific situations. Risk factors are states, demographic characteristics or events present at the time of

admission, and are dissimilar from the severity of illness scores. The risk factor based scores may be specific to a particular disease, severity, institution, therapy or technology; therefore they are not broadly generalizable.

Therapeutic Intensity based: Among the first ICU scoring systems was the Therapeutic Intervention Scoring System (TISS) [56-58]. This system is a weighted sum of the number and intensity of therapies used on a particular patient. The score is intuitively plausible, easily obtained, and extensively validated as a marker for disease intensity and costs. The disadvantage is that the discriminative power of the score rests on assumptions of appropriateness of therapy. Moreover, the therapeutic strategies vary among hospitals and physicians.

Pathophysiology based: This system assumes that regardless of disease etiology, the more disturbed a patient's physiology, the sicker or more morbid the patient. This is the organ-system-physiology approach pioneered by Knaus and associates [25] and was used in the development of Acute Physiology and Chronic Health Evaluation (APACHE) scoring system. The APACHE score is the sum of all organ system disturbances in the first 24 hours. Additional points are added for age and chronic health to achieve an overall score that has been shown to be highly correlated with mortality [59-60]. The Physiological Stability Index (PSI) [44] and Pediatric Risk of Mortality (PRISM) [61] are adapted directly from APACHE and have found widespread application in pediatric ICU research.

2.3.2 Classification of Health Status Measures by their Purpose

The health status measures can be classified by purpose into the following three categories:

A ***Discriminative measure*** focuses on cross-sectional, between-subject (inter-individual) differences on an underlying trait or dimension (e.g. anxiety, growth) when no external criterion (or gold standard) is available for validation. These measures are used to ascertain distinction between individuals or distinguish individuals from the general population, in comparative and epidemiological studies, respectively. Examples of discriminative measures are the Minnesota Multiphasic Personality Inventory (MMPI) used to compare individuals with anxiety disorder and the Alberta Infant Motor Scale used to distinguish babies with suspected developmental motor delay for monitoring and possible remedial physical therapy [62].

A ***Predictive measure*** is used to classify individuals into a set of pre-defined measurement categories when an external criterion is available, either concurrently or prospectively. Examples are the Score for Neonatal Acute Physiology (SNAP) used for predicting neonatal mortality and the Minnesota Infant Developmental Screening Test used for predicting developmental delay in infants.

An ***Evaluative measure*** focuses on longitudinal, within-subject (intra-individual) change over time on the trait or characteristic of interest. Examples include the evaluation of the

role of therapy over time on the neurobehavioral and neurosensory pattern of growth in a neonate using the Neuro-Biological Risk Score (NBRS) [38], or quantifying treatment benefit in individual subjects over time in clinical trials, e.g. 'quality of life measures' (QLM) among patients, especially in rheumatology and oncology. Occasionally a measure validated for one purpose can be effectively used for other purposes. For example, the Alberta Infant Motor Scale (AIMS) can be used for evaluating motor development in infants over time in addition to identifying infants with motor delay.

2.3.3 Implications for Development of Purpose-Specific Health Status Measures

In the past, while developing health status measures, it was not adequately stressed that defining the purpose of an outcome measure is critical for its successful design and implementation, since requirements for discriminative, predictive or evaluative measurement emphasize different, sometimes conflicting guidelines [63]. The guidelines for construction of outcome measures or indices by their purpose are illustrated in Table 2-1.

Table 2-1.

Major Purpose-Specific Issues in Construction of Health Status Measures

	DISCRIMINATIVE	PREDICTIVE	EVALUATIVE
1. Item Selection	-tap important components of the domain -universal applicability to respondents -stability over time	-statistical association with criterion measure	-tap areas related to change in health status -responsiveness to clinically significant change
2. Item Scaling	-short response sets facilitate uniform interpretation	-response sets should maximize correlations with the criterion measure	-response sets with sufficient gradations to register change
3. Basis of Item Reduction	-internal consistency -comprehensiveness -minimize standard error vs. respondent burden	-power to predict vs. respondent burden	-responsiveness vs. respondent burden
4. Validity	-structural -external-'cross-sectional'	-criterion-related	-change in individual over time -external-longitudinal
5. Reliability	-large and stable intersubject variation	-stable inter and intra-subject variation	-stable intrasubject variation
6. Responsiveness	-not relevant	-not relevant	-relevant and detects clinically important difference

This table is adapted from the work by Bram Kirshner and Gordon Guyatt at the McMaster University Health Sciences Centre [63].

2.4 Review of the Models of Structural Validation and Scaling

The structural component of validity refers to the extent to which structural relations between test items parallel what is known about the nature of the trait being measured [64]. The estimation of parameters and determination of dimensionality is essentially an effort to search for the structure in the data set, in order to gather evidence as to the structural validity of the new tool. Establishing structural validity refers to the study of the internal structure and behaviour of items in the tool. Assessments of internal consistency, standard error of measurement and comprehensiveness of the domain of the latent trait provide the basis for the reduction of items in the final version of the tool.

2.4.1 Classical Test Theory versus Item Response Theory

Much of classical test theory deals with an entire test, be it composed of one or more dimensions, and items or subsets of items. It is based on actuarial science and makes no assumptions about matters that are beyond the control of the psychometrician. These assumptions are concerned with the means and correlations of true scores and error scores. It can not predict an individual's response to items unless the items have previously been administered to similar individuals [65].

Major breakthroughs in several tool/test construction problems have been brought about through the use of latent trait models, also called item-response theory (IRT) models [66-69]. The IRT is gaining in acceptance in medical, psychological and educational research because it provides more adaptable and effective methods of tool construction, analysis

and scoring than those derived from the classical test theory. These probabilistic models are based on stronger assumptions about the data. The IRT is based on the notion that a subject's position on the latent continuum determines his performance on items measuring that trait. The response models on which IRT models are based enable the analyst to estimate the probability that a respondent at a particular score/scale level will endorse a given item. This permits 'content referencing' of the scale scores, i.e. the indicator items become milestones on the continuum.

The property of invariance of item and ability parameters is the cornerstone of IRT and its major distinction from the classical test theory. This property implies that the parameters that characterize an item do not depend on the latent-trait (e.g. morbidity) distribution of the study subjects and the parameters that characterize a subject do not depend on the set of items provided in the test [65,68,70]. The source of IRT's greater power is in the relationships it establishes between the properties of the items and the operating characteristics of the tool made up of the items. A unique property to IRT is the location of the items and the subjects on the same scale. In latent trait models, the *difficulty parameter estimate* of items is accounted for by the model and reflects in the *ability parameter estimate* of subjects. Yet another desirable property of the IRT models is that they provide a measure of the precision of trait estimation at each trait level. Thus, instead of providing a single standard error of measurement that applies to all respondents irrespective of their trait level, IRT models make it possible to provide separate estimates of error for each trait level or each respondent.

2.4.2 Item Response Theory Model

A typical structural validation procedure using an IRT model involves the following steps including configuration of the model for dimensionality, item analysis and item parameter estimation, evaluation of the degree of fit of items on the scale, and recommendation of scaling and scoring models.

2.4.2.1 Configuration of the Model for Dimensionality

Determination of the number of dimensions underlying a data set is a central problem in measurement, and one that must be addressed prior to the application of the specific methods for parameter estimation and scaling of items. Dimensionality is a property of both the measure and the study population being measured. McDonald in 1985 defined a unidimensional test as a test whose items fit a latent trait or common factor model, possibly non-linear, with just one latent trait or common factor [72]. In practice, the dimensionality is situation-specific, i.e. dimensionality is not a property of the items but rather of the responses to items under a specified set of conditions.

Factor analysis has long been used to test dimensionality of a data set. Based on correlational procedures, factor analysis attempts to account for the observed correlations by extracting a set of factors and determining dimensionality from the correlation matrix. More recent research has shown that factor analysis is not the most sensitive procedure for detecting multidimensionality in the context of IRT [59,61,62]. Stout proposed a non-parametric statistical test of unidimensionality based on the concept of essential

unidimensionality for binary IRT [77]. The approach attempts to distinguish dominant dimensions in a test, from minor or nuisance dimensions. It provides a statistical test based on a large sample distribution theory, for assessing latent trait dimensionality [77,78].

Stout's test of dimensionality, known as DIMTEST has been shown to discriminate well between one and more dimensional tests, maintaining good adherence to a specified level of significance with unidimensional measures and good power with more dimensions. For these reasons we elected to use the DIMTEST for testing the unidimensionality of our data set before undertaking parameter estimation and calibration using the IRT.

The procedure for determining the dimensionality using DIMTEST is approached in three steps. The first orientation is to assess the lack of unidimensionality of a test data set. To achieve this, either exploratory statistical procedures or content-knowledge application can be used to select a subset of items. The second orientation assesses whether the particular specified subset of items is dimensionally distinct from the remainder of the test. The third orientation determines the dimensionality structure of the test. It is analogous to the multiple regression problem of selecting a set of influential independent variable from a large initial set of variables. It is useful to carry out multiple iterations of DIMTEST, each time for a distinct subset of items. The basic underlying principle of DIMTEST is that if unidimensionality holds then the basic IRT assumptions of local independence holds approximately within each subgroup, and hence the two within-subgroup variance estimates should be nearly equal.

Items in a tool are said to be multiply determined, that is in addition to measuring the intended attribute, other attributes unique to individual items or common to a relatively few items are unavoidable. An example of measuring more than one dimension in an item in our data set is *need for mechanical ventilation beyond 7 days of age*, where the morbidity attribute is confounded with therapeutic management.

In this study, the DIMTEST technique tested the hypothesis that an essentially unidimensional latent trait model fits the observed binary item response data in the data set. The statistical procedure [79, 80] for testing the null hypothesis of essential unidimensionality of items in the NMI consists of the following steps:

(a) The N index binary items are split into two assessment subsets of length M each, called the Assessment I subset (AT1) and the Assessment 2 subset (AT2), and a longer subset, called the partitioning subset (PT) of length n (=N-2M).

(b) The M items for subset AT1 are selected to have the same dominant trait. This selection can be done using either content knowledge or exploratory factor analysis. In either case, the goal is to select a small subset of items (preferably up to one fourth of the total number of items) that measure the same dominant trait and are as dimensionally different as possible from the PT items.

(c) A second set of M items for AT2 is selected from the remaining items so that AT2 items have a difficulty distribution similar to AT1 items.

(d) The remaining n (=N-2M) items then become the partitioning subset PT.

(e) Each subject is assigned to one of the K subgroups according to the subject's score on the PT.

After eliminating subgroups with too few subjects (J=20 recommended), within each subgroup K, two variance estimates, the usual variance estimate and the unidimensional variance estimate are computed using items of AT1.

(f) The difference in these variance estimates then is normalized and summed over subgroups to arrive at a statistic, T_L.

(g) Similarly, using items of AT2, statistic T_B is calculated.

(h) The statistic T, to assess departure from essential unidimensionality, is given by

$$T=(T_L-T_B)/\sqrt{2}$$

(i) The null hypothesis is that the test is unidimensional. It is rejected if $T \geq Z_\alpha$, where Z_α is the upper 100 $(1-\alpha)$ percentile of the standard distribution and α is the desired level of significance.

2.4.2.2 Item Analysis and Parameter Estimation using Item Response Theory

The process of estimation of the relationship between item performance and the underlying trait (e.g. newborn morbidity) is known as parameter estimation. The principal features of the IRT model considered for item analysis and parameter estimation are as follows:

IRT Models by Parametric Structure: Within IRT, the models are distinguished by the shape of the item characteristic curves (i.e., normal ogive or logistic) and by the number of parameters being estimated. The two most salient choices are between normal-ogive and logistic versions of models. The two are very similar to each other. However,

logistic models are mathematically convenient and are more robust to unwarranted fluctuations at asymptotes, i.e. pseudo-chance errors.

The three most popular unidimensional IRT models for parameter estimation are the *one, two* and *three-parameter* logistic models, so named because of the number of item parameters each incorporates. The choice involves assumptions about the data that can be verified later by examining the goodness-of-fit [81]. Models in which only item-locations, b_i (difficulties), are of concern are referred to as one-parameter logistic models or Rasch models. In the one-parameter model, it is assumed that b_i is the only item characteristic that influences a subject's performance on the latent-trait scale (hereby called morbidity scale). The two-parameter model has an additional element, a_i, which is called the item discrimination parameter. Therefore, two-parameter models are those concerned with both the item-location (b_i) and the discrimination (a_i) parameters. The third parameter, c_i, in three-parameter models is incorporated to take into account performance at the low end of the continuum. It represents the probability of subject with low latent-trait endorsing on a high trait item by guessing (chance), which is an unlikely situation in the case of the proposed morbidity scale. Therefore, c_i, the guessing or pseudo-chance parameter is irrelevant for our outcome measure.

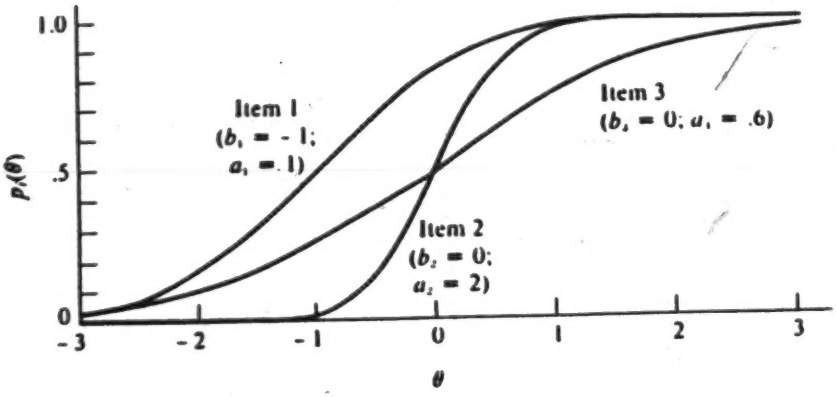
The mathematical framework of IRT with its set of assumptions and statistical constraints has been presumed to provide an adequate description of the data in our study. The three assumptions underlying the commonly used IRT models are *unidimensionality, local independence* and *monotonicity*. Unidimensionality assumes that the items of a test measure a single underlying latent trait. A respondent's position in the latent space is

determined from his/her latent trait scores. The dimensionality of the latent space depends on the number of traits that underlie performance on the test items. Local independence assumes that given any particular level of the latent-trait (morbidity), responses to different items are independent of each other i.e. only the trait accounts for systematic differences between groups. When two items are locally independent, the probability that a subject with a given latent trait value endorses both items is the product of the probabilities that the subject endorses each item. Local independence represents a restrictive assumption and dimensionality testing including factor analytic techniques can test its appropriateness. In these probabilistic models, it is assumed that the probability of endorsing an item increases monotonically with the level of the underlying trait. An item characteristic curve (ICC) so produced is a mathematical function that relates the probability of endorsing an item to the trait level measured by the item set or total index. It is a non-linear regression function of item score on the latent trait measured by the test. The item characteristic curve is defined completely when its general form (each curve is a member of a family of curves of the general form as discussed by Lord and Novick [68] and Togerson (1958) [71]) is specified and its parameters, denoted a_i and b_i, are known. Referring to Figure 2.1, the a_i parameter is proportional to the slope of the ICC at point b_i, the item's location on the latent-trait scale. Parameter b_i is defined such that a subject has a 50% chance of endorsing an item i when $\theta = b_i$ (i.e. the subject's trait level matches the item difficulty parameter). The latent-trait at which an item discriminates most effectively among subjects is b_i, and the ICC is steepest at b_i. The slope of the curve at b_i is equivalent to parameter a_i. Items with steeper slopes are more useful for separating newborns into different trait (morbidity) levels than are items with less steep slopes.

Figure 2.1.

Item Characteristic Curve

[b = intercept; a = slope]



For example, item 1 discriminates best at θ=-1 and also gives some information about θ anywhere between -2 and 0. Item 2 discriminates very well near 0 and gives little information about other trait levels, its information curve is fairly flat for θ less than -0.5 or greater than 0.5. Item 3 also reflects θ best near 0, but gives some more information about θ for values of θ from -2 to +2. Item 3 is less discriminating than item 2 at θ=0.

Birnbaum in 1968 [82] proposed a latent trait model in which the item characteristic curve takes the form of a two-parameter logistic distribution function given by the following mathematical form:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (i=1,2,3....n).$$

In this equation $P_i(\theta)$ is the probability that a subject with the trait value θ endorse item i,

a_i and b_i are parameters for item i (i=1,2...n) and n is the number of items in the test. The constant D (D=1.7) is the scaling factor for a logistic model. An additional implicit assumption of this model is that the pseudo-chance parameter c_i is kept equal to zero. In the NMI, the latent trait value, θ, represents the level of morbidity in a newborn, while the item location parameter, b_i, indicates the position of that item on the morbidity continuum. The discrimination parameter, a_i, is an indication of the point of maximum information and the discriminatory capability of an item on the morbidity continuum. An item response theory for a two-parameter model postulates that the newborn's level of morbidity together with the characteristics, a_i and b_i of morbidity items accounts for the newborn's behaviour on an item.

A special case of the Birnbaum's two-parameter logistic model, in which all the items are assumed to have equal discriminating power, and vary only in terms of the difficulty parameter is given by Rasch [81]. The equation of the ICC for this one parameter model is written as:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (I = 1,2,3....n);$$

where a_i, is the common level of discrimination for all the items and is held constant.

Response Level: Three types of response levels for data are described, i.e. dichotomous, polychotomous, and continuous. Many of the morbidity items in the present study have several categories and the number of categories of the individual

attributes of morbidity vary across the item-list (refer Appendix 7), (e.g. the attribute *cord blood pH* has two categories, whereas the attribute *flaccidity* has only one, and the attribute *seizures* has four categories). For this reason, we coded the items in a dichotomous (yes/no) fashion, to be treated singly as *units of information*, to enable analysis using BILOG [83] rather than MULTILOG software program. With the program MULTILOG, analysis is performed using attributes with equal number of categories in each attribute, which was not the case in our data set. This particular fashion of data handling to enable the use of BILOG necessitated a careful examination of the assumption of *local independence* in the present data set. Further analysis revealed that the benefits of the model using BILOG might overcome the minor violations of the assumption of *local independence*.

Item Calibration: Estimating the item parameters and checking the fit of the models is referred to as item calibration. To make use of the IRT for test scoring, the parameters for each item of the index must be estimated first using these models.

The morbidity scale can be defined as a scale on which ICCs have some specified mathematical form. The origin and unit of measurement of the ability scores are arbitrary. The morbidity scale can be stretched and compressed at different locations so as to maximize the fit among item responses, ICCs and morbidity scores. Using appropriate significance tests, the fit of the model to the test data can be assessed with the help of appropriate computer software.

In the present study, the BILOG [83] was used to test an Item Response Theory model and calibrate items in the proposed index. The BILOG program is designed for a wide range of applications of IRT to practical tool development problems. It assumes binary (right-wrong) scoring of items, and gives stable and accurate estimates of item parameters and scale scores for both long and short indices.

Following are some of the special features of the BILOG program:

(a) Choice of 1, 2, or 3-parameter logistic item response models

(b) Marginal maximum likelihood estimation of item intercepts, slopes, and lower asymptotes

(c) Standardized posterior residuals for individual items

(d) Tests of fit for individual items

(e) Standard errors for all item parameter estimates and scale scores of respondents

(f) Several options for handling omitted items

(g) Test and item information analysis

(h) Plots of item-response functions, item information curves, and test information curves

(i) Rescaling of test scores

(j) Estimation of latent distribution

(k) Employs a conditioned Newton-Raphson procedure to estimate item parameters.

(l) Chi-square item fit statistics is calculated when the number of items is ≥ 20.

Parameter Estimation Techniques: The literature of latent trait theory abounds with procedures for estimating the parameters that arise in such models. Three common types of scale score estimation techniques in general use are Marginal Maximum

Likelihood Estimation (MLE), Bayes Estimation, and Marginal Maximum A Posteriori Estimation (MMAP). A robust form of estimation, MLE technique, using software BILOG, performed the parameter estimation of the NMI. This approach applies to all types of item response models and can be distinguished into conditional and unconditional, the former being more common of the two. The conditional method assumes the independence of responses to different items by persons of the same morbidity θ. Because the joint probability of independent events is the product of the probabilities of the separate events, this assumption makes it possible to calculate the probability of observing a particular pattern of item scores, in the responses of a person with morbidity θ. The value of θ that makes the likelihood function for a subject a maximum is defined as the maximum likelihood estimate of θ for that subject. The so-called "EM" algorithm and Newton-Gauss (Fisher scoring) methods are used to solve the likelihood equations [83]. Details of these may be found in Bock and Anderson [84-87]. Standard errors and correlations of the parameter estimators are obtained by inverting the information matrix in the Fisher-scoring solution. It is also possible to estimate the distribution of θ by MLE. A widely accepted convention to solve the indeterminacy of location and scale inherent in these solutions is by setting the mean of the latent distribution of θ to 0 and the standard deviation of the distribution to 1. A disadvantage of the MLE estimation is that it is not defined for the response patterns in which all or none of the items are endorsed.

Test Information Function: Once an IRT model is specified, the precision with which it estimates a subject's latent trait (morbidity) can be determined using the test

information function and efficiency of the estimated parameters. The test information consists entirely of independent and additive contributions from the items. The contribution of an item does not depend on what other items are included in the test. The contribution of a single item is called the item information. The test information function is inversely proportional to the squared length of the standard error for estimating morbidity (latent trait) from test scores.

The information provided by score y for estimating θ varies at different θ levels. The variation is due to two sources: (a) the standard error of estimation (SEE), the smaller the SEE, the more information y provides about θ, (b) the steeper the slope of the regression, the more information y provides about θ. The standard error is given by the formula:

$$SE(\theta) = 1/I(\theta)^{1\backslash 2}$$

where, $I(\theta)$ is called the information function.

The normality of the estimated θ can be used to construct a confidence interval for θ. The $(1-\alpha)$ % confidence interval for θ is given by

$$[(\text{estimated } \theta) - Z\alpha/2\ SE(\theta),\ (\text{estimated } \theta) + Z\alpha/2\ SE(\theta)]$$

where, $Z\alpha/2$ is the upper $(1-\alpha/2)$ percentile point of the normal distribution. For the 95% confidence interval (CI), $\alpha = 0.05$ and $Z\alpha/2 = 1.96$. Therefore, when information at a morbidity level is high, narrow confidence bands are computed around the estimates and vice versa for low information. It is important to note that the value of $SE(\theta)$ varies with morbidity level.

2.4.2.3 Evaluation of the Degree of Fit of Items on the Scale

Item response models offer a number of advantages for test score interpretations and reporting of test results, but the advantages will be obtained only when there is a close match between the model selected for use and the test data. The determination of how well an item response model fits a set of test data is addressed by both statistical and non-statistical considerations of the goodness-of-fit model [88]. The following implications were carefully examined in regard to testing the assumptions and the results of analyses are discussed in Chapter 4.

(a) Whether the test data satisfied the assumptions of the test model of interest:

Model selection was aided by the investigations of

(i) Dimensionality testing

(ii) Plot of content-based versus test-based item parameter estimates

(b) Whether the expected advantages derived from the use of the IRT model were obtained:

The two expected advantages are the invariant estimates of *item* and *morbidity-trait.*

(c) Whether there is a closeness of fit between predictions and observable outcomes:

This was determined using the following tests:

(i) Chi-Square

(ii) Standardized posterior residuals

(iii) Plots of test scores and latent-trait estimates

2.4.2.4 Recommendation of Scaling and Scoring Models

The recommendations are based on the following principles:

(a) Properties of the Final Scale

Togerson [71] makes the following points regarding the properties of the scale. If an attribute is to be represented numerically, an isomorphism or one-to-one relationship must exist between the characteristics of the number system and the relations between the quantities of the attribute to be measured. The formal number system possesses the properties of order (numbers are ordered), distances (differences between numbers are ordered), and origin (the zero point). Order is involved in all scaling methods, so the types of scales achieved are distinguished by which of the other properties they possess. This gives rise to four types of scales: a) ordinal, b) ordinal with natural origin, c) interval, and d) ratio. In this study, the properties of order and distance seem relevant to the task of scaling this set of morbidity items. Therefore, the final scale would provide scale values with the interval level properties derived from the ordinal level data.

(b) Determination of the Scoring System

The recommendation of a model for scoring is based on the results of structure found by dimensionality analysis and parameter estimation of the latent trait under consideration. Three scoring systems named additive, configured and milestone were considered for appropriateness to the proposed index. The additive model locates an infant on the scale by counting the number of items endorsed. In the configured model, the patterns of

responses are used for scoring, while the highest item(s) achieved is scored in the case of the milestone model. We further explored the additive and milestone scoring models, the latter aggregated the highest three or five items scored by a newborn on the morbidity scale.

Chapter 3

Subjects and Methods

The research project consisted of three major components, *validation, scaling and scoring* of the proposed outcome measure, the Newborn Morbidity Index. The validation process consisted of three phases, substantive, structural (including scaling) and external. The procedures and methods of data collection for the study are described in phases corresponding to the validation process. Definitions of the frequently used measurement terms are provided in Appendix 1.

3.1 Substantive Validation

The research plan for substantive validation included the analysis of literature, item generation, item review, item selection and preliminary item testing. This was accomplished between January-September 1995, and the procedures for substantive validation are described in the following three steps:

3.1.1 Analysis of Literature on the Structure of Newborn Morbidity

A literature search using MEDLINE and HealthSTAR yielded a list of approximately thirty inventories or indices broadly used for various different purposes in the field of maternal-child health [29-43,46]. A thorough review of these indices provided guidelines for the specification of the domain of the construct, newborn morbidity, its manifest

items, and desirable features, relevant to the characteristics of our proposed outcome measure.

We focused on the following five features of manifest items that conformed to the construct theory of newborn morbidity:

(a) Relevance: Relevance was determined by literature review and consultations with the subject-matter experts including five neonatologists and pediatricians, and three obstetricians from the University of Alberta. Included manifest items were relevant to the proposed purpose of the outcome measure as well as to the study population. Important relevant characteristics of the selected items were being predominantly pathophysiology-item based, and simple to use.

(b) Representativeness: A review of the available literature [20-23] showed that the observable indicators that reflect morbidity in newborns at birth are the attributes of vital signs (such as heart rate, respiratory rate, blood pressure, color and temperature), clinical features due to disturbed organ systems, and the results of biochemical and physiological investigations that evaluate the disease state.

An attempt was made to obtain an adequate sampling of the domain of acute pathophysiology from across the entire continuum of morbidity. Conforming to existing views on newborn morbidity in the literature, morbidity was represented by the manifest items of clinical features, vital signs and a few biochemical tests in newborns, up to

seven completed days after birth. We specified the domain of *manifest items* of morbidity by grouping them into four categories according to the time of their onset, i.e. at birth, within 24 hours of birth, up to seven days and persisting beyond seven days after birth (before discharge from the hospital). Table 3-1 contains a list of the selected manifest items grouped into four categories.

Table 3-1.

Table of Specifications of the Selected Manifest Items of Newborn Morbidity Items Listed by the Time of Assessment

At Birth	Worst Physiology Within 24 hrs. of birth	Up to day 7	Persisting beyond day 7
Apgar Score	Heart Rate	Seizures	Respiratory Distress
Cord Blood pH	Respiratory Rate	Urine Output Bleeding Disorder	Oral Feeding Difficulties
Meconium	Systolic B.P.	Hypotonia & Flaccidity Apnea	Level of (not induced)
Resuscitation at Birth	Colour	Hypoglycemia Hyperbilirubinemia	Cardio-Pulmonary Resuscitation Intra-ventricular Hemorrhage
Trauma		Bacterial Culture	

(c) Internal Consistency: Internal consistency means that the selected items were affected in the same or related ways in response to various different maternal conditions, settings or obstetrical situations. Cronbach's alpha or reliability coefficient is a measure of internal consistency of a tool. The results of analysis of this measure are discussed in chapter 4.

(d) Standard Assessments: The selected manifest items signified standard assessments done by the health care providers. Items that required special settings, resources or elaborate laboratory testing were avoided so as to promote generalizability of the outcome measure to various obstetrical settings and situations.

Our set of manifest items includes a limited number of routine biochemical or physiological diagnostic tests, since the inclusion of a large number of laboratory tests will likely result in a tool that focuses on dimensions (or constructs) other than morbidity. Additionally, the resulting index would be dependent on the context of measurement of those diagnostic tests.

(e) Usability: The selected manifest items represented morbidity of interest relevant to most maternal diseases and conditions, thereby promoting usability of the outcome measure to most obstetrical trials.

Following this analysis, a provisional set of manifest items including the attributes of acute pathophysiology and possessing the specified desirable features was compiled. The items in the provisional set were grouped into four categories by the time of their assessment, as given in Table 3-1. The timing of assessment of morbid events bears clinical relevance to the progression of morbidity. For example, a newborn with respiratory distress in the first 3 hours of birth is less morbid than a newborn with respiratory distress in the first 24 hours. Similarly, the latter is less morbid than a newborn with respiratory distress lasting for 7 days.

3.1.2 Item Generation and Item Review

For the purpose of item-generation in the provisional set of items, we employed as a prototype, the index of perinatal morbidity and mortality developed by Hannah et al. [46] in 1985 to serve as a discriminative measure of perinatal outcome. During their trial studying the management of post-term pregnancies, Hannah et al. compiled an item-list by adjudication of all potentially abnormal newborn morbidity items among 1500 post-term infants enrolled. A panel of experts from Neonatology and Obstetrics was responsible for their item-generation and weighting of items. There were 22 items in their item-list, including *mortality* and *length of stay* in the NICU. Although the items *mortality* and *length of stay* are reflections of morbid conditions, they are external to the domain of morbidity. Therefore, they were excluded from our item-list. Item *Apgar score at 1 minute* was excluded, while items *Apgar score at 5* and *10 minutes* were retained, because the latter Apgar scores are better reflections of lasting significant morbidity [9-11].

Subsequently, we contacted the subject-matter experts at the University of Alberta including five neonatologists, one developmental pediatrician and three obstetricians, and informed them about the purpose of our study. On recommendations by the subject-matter experts during a preliminary working session, seven items were added to the provisional list. These items included vital signs (color, blood pressure, heart rate), urine output, bleeding disorder, bacterial culture, and serum glucose level. Following these modifications, we compiled a revised list of attributes of morbidity possessing the

specified desirable characteristics of the manifest items. Each morbidity attribute was explicitly defined and sub-divided into categories by the levels of severity.

A working session was held with ten reviewers consisting of subject-matter experts (seven neonatologists) and end-users (three obstetricians). The purpose of this session was to ascertain the relevance and representativeness of the included items, and to ensure their clarity and accuracy. An information sheet was sent to the subject-matter experts for the assessment of the items of morbidity, and is included as Appendix 2. The working session was comprised of three stages spread over seven days. First, the experts were each given a copy of the item sets and were asked to review items with respect to clarity, inclusiveness, significance and level of severity. Second, suggestions were sought regarding modification, deletion or refinement of the items. The third stage involved rating of each item on an analogue scale of 0 to 10, where 0 was equivalent to 'no morbidity' and 10 to the 'most severe' morbidity. Responses on the rating scale were obtained from six reviewers including three subject-matter experts and three end-users. This rating of items relates to substantive validity and reflects the experts' opinions that the levels-of-severity were unequivocal and the items were aligned on a continuum of morbidity. The average ratings for the items along the continuum of morbidity are shown in Table 3-2, in an ascending order by their average weight. At the end of this session, a refined, revised list of twenty-two attributes of morbidity was produced. The categories of the attributes of morbidity were coded as yes/no, in a binary fashion. There were a total of sixty binary items in the item-list at the time of rating by the experts.

Table 3-2.

Morbidity Items in Ascending Order by Average Weight
Number of items=60

Serial No.	Morbidity Item	Ave. Wt.	SD	BIN
1	Meconium above cords	0.40	0.22	5
2	Cephalhematoma	0.45	0.37	49
3	Hypotonia identified at <1 hr.	0.70	0.28	22
4	Serum bilirubin >170 micromol/L	0.80	0.24	60
5	Resuscitation at birth (with bag & mask)	0.90	0.22	3
6	Apgar score of <7 at 5 minute	0.90	0.22	7
7	Apnea detected (by apnea monitor)	0.90	0.42	26
8	Jitteriness/Tremors	0.90	0.22	31
9	Heart rate/minute (160-200)	1.08	0.66	16
10	Hyperalert/Hypertonic	1.20	0.26	35
11	Blood glucose >1.7 mmol/L and <2.2 mmol/L	1.25	0.62	58
12	Cord blood pH <7.10	1.60	0.42	1
13	Poor sucking within 24 hrs.	1.60	0.42	39
14	Respiratory rate/minute <30 or >60 in first 24 hrs. (2 or more	1.70	0.28	17
15	Heart rate/minute >200 in 24 hrs. (2 or more consecutive rea	1.75	0.26	14
16	Serum bilirubin - >250 micromol/L OR Need for phototherapy	1.90	0.74	61
17	Apnea and need for oxygen therapy	2.00	0.36	27
18	Meconium below cords	2.25	0.26	6
19	Apgar score <7 at 10 minute	2.40	0.55	10
20	Respiratory rate/minute >100 in 3-24 hours	2.50	0.50	18
21	Flaccidity between 0-120 hrs.	2.50	1.00	25
22	Heart rate/minute <100/minute (2 or more consecutive readi	2.58	0.49	15
23	Assisted ventilation within 24 hrs.	2.60	0.41	43
24	Persistent vomiting	2.67	0.55	42
25	Low systolic BP	2.67	0.26	16
26	Low urine output (<2 ml/kg/hour)	2.70	0.45	48
27	Apgar score <4 at 5 minute	2.70	0.45	8
28	Resuscitation at birth (with intubation)	2.83	0.26	4
29	Fracture long bone/clavicle	2.92	0.49	50
30	Cord blood pH <7.0	2.92	0.80	2
31	Level of consciousness- drowsy/lethargic	3.0	0.78	36
32	Poor sucking between 24 hrs.- day 7	3.0	0.89	40

Table 3-2.

Morbidity Items in Ascending Order by Average Weight
Number of items=60

Serial No.	Morbidity Item	Ave. Wt.	SD	BIN
33	Plasma glucose (<1.7 mmol/L)	3.08	1.07	59
34	Hypotonia between 1-120 hrs.	3.08	1.00	23
35	Thrombocytopenia with or without bleeding disorder	3.10	0.75	29
36	Altered colour - dusky/central cyanosis	3.25	0.53	21
37	Bacterial culture positive-blood	3.25	0.69	63
38	Single seizure	3.30	0.67	32
39	Serum bilirubin >340 mmol/L OR Exchange transfusion	3.30	0.21	62
40	Apnea and need for resuscitation therapy	3.50	0.50	28
41	Assisted ventilation beyond 24 hrs.	3.75	0.61	44
42	Bleeding disorder- need for transfusion due to item 35	4.17	0.75	30
43	Birth trauma-nerve injury (facial/peripheral) - resolved at discl	4.33	0.41	54
44	Birth trauma-nerve injury (facial/peripheral) - not resolved	5.2	0.57	55
45	Poor sucking beyond day 7	4.42	0.92	41
46	Multiple seizures	4.60	0.55	33
47	Hypotonia beyond 120 hrs.	4.80	0.84	24
48	Mechanical ventilation within 24 hrs.	4.80	0.57	45
49	Apgar score <4 at 10 minute	4.83	0.75	11
50	Apgar score <1 at 5 minute	5.17	0.41	9
51	Mechanical ventilation between 24 hrs.- 7 days	5.50	0.49	46
52	Subdural hematoma	5.58	1.17	55
53	Seizures and >2 drugs used for treatment	6.00	0.36	34
54	Cardio-pulmonary resuscitation any time before discharge	6.08	0.73	38
55	Bacterial culture positive - CSF	6.67	0.82	64
56	Intracerebral hematoma	7.00	1.26	56
57	Need for mechanical ventilation beyond day 7	7.42	1.02	47
58	Level of consciousness- stupor/obtundation/coma	8.20	0.84	37
59	Spinal cord injury	8.5	0.55	57
60	Apgar score <1 at 10 minute	9.33	1.03	12

Serial No.= Serial number for items in this table; Ave. Wt. - Average of the ratings by the six content-area experts; SD= Standard Deviation; BIN = Binary Item Number (refer page 45, chapter 3 for the definition)

3.1.3 Preliminary Item Testing

The researcher conducted a preliminary testing of the items contained in the revised item-list. This testing was carried out by examining the charts at the Department of Health Records of one hundred newborns discharged from the NICU. The purpose of this testing was to ensure precision, accuracy and extractability of the charted information on each item included in the item-list. Subsequent to the preliminary testing, categories on two of the attributes in the item-list were further revised to conform to the pattern of recording information in the newborn's flow sheet. Those attributes were *altered colour* and *level of consciousness*. The categories of the attribute *altered colour* (Serial No. I; Table 3-3) were changed from *peripheral cyanosis* and *central cyanosis* to include *ruddy* and *dusky*, respectively. The categories of the attribute *level of consciousness* (Serial No. O; Table 3-3) were revised to include *hypertonia* with *hyperalert* and *hyperirritable* state. The attributes of *flaccidity* and *intra-ventricular hemorrhage* were included after consultations with the experts.

The revised, substantively-validated morbidity item-list consisted of twenty-four attributes of morbidity which were further subdivided into categories by the level-of-attribute (severity of morbidity) approach, and is shown in Table 3-3. The first column in Table 3-3 contains the serial number (S.No.) of the morbidity items. For item-analysis, the items were coded as binary digits (yes/no or 1/0) and were sixty-six in number in this (provisional) morbidity item-list. The column labeled BIN contains the binary item number assigned to the morbidity item.

Table 3-3.

Provisional Morbidity Item-List
Number of Attributes of Morbidity = 24
Number of Binary Items = 66

SNo.	Morbidity Item	BIN
A	Cord Blood pH	
a	<=7.1	1
b	<7.0	2
B	Resuscitation at Birth	
a	Bag and mask	3
b	Intubation	4
C	Meconium	
a	Meconium above cords	5
b	Meconium above and below cords	6
D	Apgar Score (5 minute)	
a	score<7	7
b	score<4	8
c	score<1	9
E	Apgar Score (10 minute)	
a	score<7	10
b	score<4	11
c	score<1	12
F	Heart Rate/minute	
a	161-200	13
b	>200	14
c	<100	15
G	Systolic BP (mean, mm of Hg)	
	28-32 weeks 32-42 weeks	
a	<30 <40	16
H	Respiratory Rate/minute	
	for >2 consecutive readings	
a	<30 or >60 in the first 3 hrs.	17
b	>100 between 3-24 hrs.	18
c	<30 or >60 between 3-24 hrs.	19

SNo.	Morbidity Item	BIN
I	Altered Colour	
a	Ruddy OR Peripheral cyanosis	20
b	Dusky OR Central cyanosis	21
J	Hypotonia	
a	Identified at <1 hr.	22
b	Persisting (or Identified) after 1 hr. of age	23
c	Persisting after 120 hrs. of age	24
K	Flaccidity	
a	Present (between 1-120 hrs.)	25
L	Apnea	
a	Apnea detected (by apnea monitor)	26
b	Apnea and need for oxygen	27
c	Apnea and need for resuscitation	28
M	Bleeding Disorder	
a	Thrombocytopenia with or without	29
	bleeding disorder (GI OR Lungs OR Skin)	
b	Need for transfusion due to M-a	30
N	Seizures	
a	Jittery OR Tremors	31
b	Single seizure	32
c	>Single seizure (Multiple seizures)	33
d	If >2 drugs used for treatment in N-c	34
O	Level of Consciousness	
a	Hyperalert OR Hypertonic	35
b	Drowsy OR Lethargic	36
c	Stupor OR Obtundation OR Coma	37
P	Cardio-Pulmonary Resuscitation	
a	Any time before discharge	38

Table 3-3.

Provisional Morbidity Item-List
Number of Attributes of Morbidity = 24
Number of Binary Items = 66

SNo.	Morbidity Item	BIN	SNo.	Morbidity Item	BIN
Q	**Oral Feeding Difficulties**		U	**Hypoglycemia (lowest level)**	
a	Poor sucking within 24 hrs.	39	*a*	Blood glucose <2.2 mmol/l	58
b	Poor sucking between 24 hrs.-7 days	40	*b*	Blood glucose <1.7 mmol/l	59
c	Poor sucking beyond day 7	41	V	**Hyperbilirubinemia, micromol /L (peak level)**	
d	Persistent vomiting	42	*a*	Serum bilirubin >170	60
R	**Respiratory Status**		*b*	Serum bilirubin >250 OR Phototherapy	61
a	Assisted ventilation within 24 hrs.	43	*c*	Serum bilirubin >340 OR Exchange transfusio	62
b	Assisted ventilation beyond 24 hrs.	44	W	**Bacterial Culture**	
c	Mechanical ventilation within 24 hrs.	45	*a*	Blood positive	63
d	Mechanical ventilation 24 hrs.-7 days	46	*b*	CSF positive	64
e	Mechanical ventilation beyond day 7	47	X	**Intra-ventricular Hemorrhage**	
S	**Urine Output**		*a*	Grade 1 or 2	65
a	Low (< 2ml/ kg/ hour)	48	*b*	Grade 3 or 4	66
T	**Birth Trauma**				
a	Cephalhematoma	49			
b	Fracture of long bone OR clavicle	50			
c	Fracture of skull	51			
	Nerve injury (facial OR peripheral)				
d	detected and resolved on day 1	52			
e	resolved after day 1	53			
f	not resolved (at discharge)	54			
g	Subdural hematoma	55			
h	Intracerebral hematoma	56			
i	Spinal cord injury	57			

SNo. = Serial Number for items; BIN = Binary Item Number (refer page 45, Chapter 3)

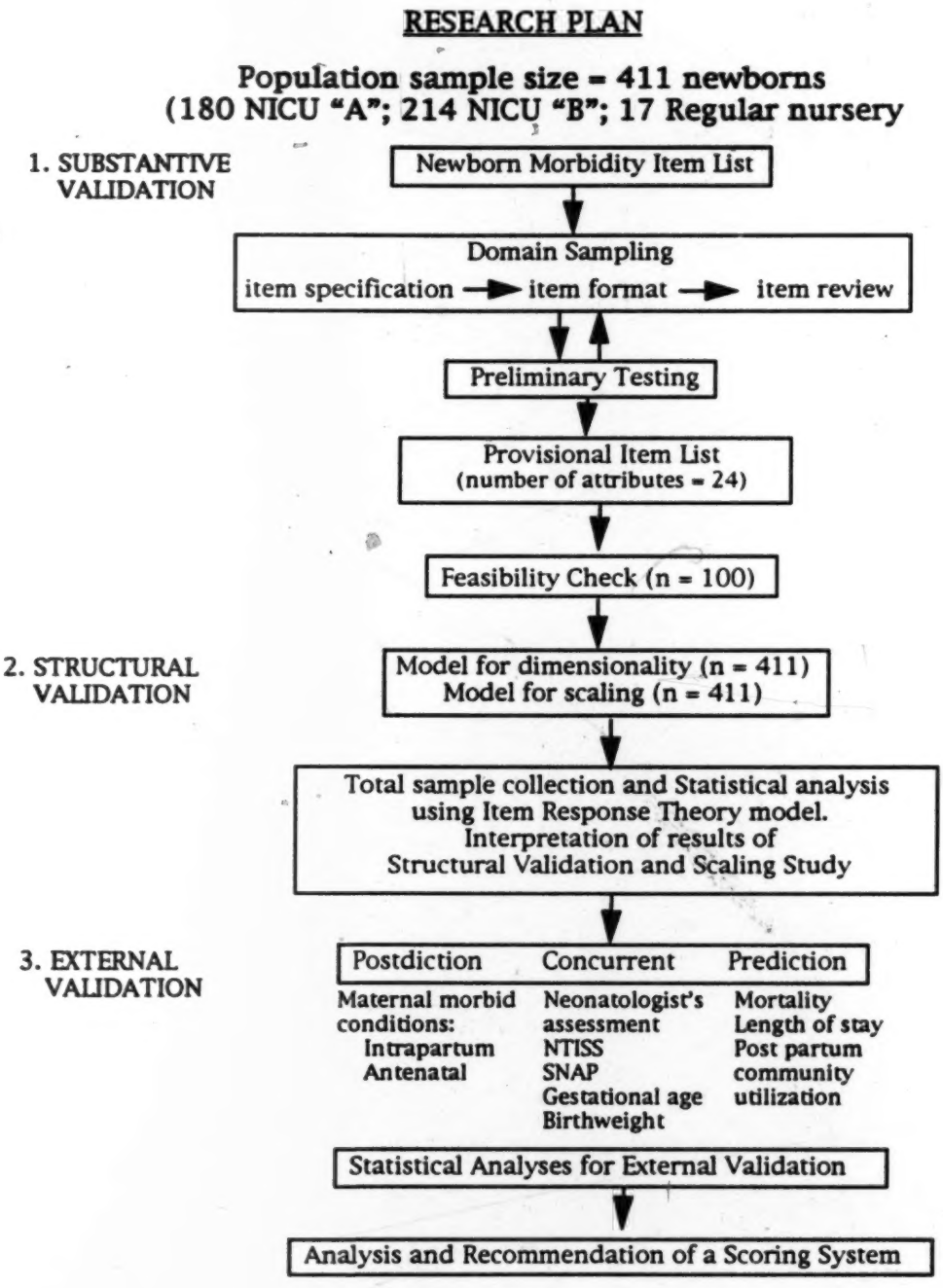
Where applicable, a continuity in the underlying morbidity trait was assumed across various categories by the level of severity of an attribute; such that a newborn with a cord blood pH of 6.9 would score on both category A.a and A.b. Similarly, a newborn with an Apgar score of 0 at 5 minutes would score on each of the categories, D.a, D.b and D.c of the attribute D.

The included attributes of morbidity vary in the number of categories across the item-list. For example, the attribute *cord blood pH* has two categories, whereas the attribute *flaccidity* has only one and the attribute *seizures* has four categories. Because of the unequal number of categories in each item, the included binary items will be analyzed using the BILOG [81] rather than MULTILOG program, where analysis is performed using attributes with equal number of categories. The subsequent phase of validation will determine the dimensionality of the latent trait [77-79] to resolve the issue of local dependence among binary items. Local dependence can be attributed to dependence between any two or more morbidity attributes, or that arising due to the assumption of continuity of the underlying trait, as discussed above.

Summaries of the modifications to the item-list at each step of the substantive validation are provided in Table 4-5 of Chapter 4. The substantive validation resulted in an item-list consisting of sixty-six binary items belonging to twenty-four attributes of morbidity as also illustrated in Table 3-3. This item-list was used for data collection on newborn morbidity items in the subsequent phases of the study.

3.2 Structural Validation and Scaling

The adjoining flow-chart illustrates the research plan and design for the study.



3.2.1 Study Site and Sample Size

The study sample for the structural and external validation phases of the project was recruited at the Royal Alexandra Hospital (RAH) in Edmonton, Alberta. The hospital is equipped with a tertiary-care neonatal intensive care unit (NICU) and approximately 5,500 women deliver at this facility per year. An estimated 1200 newborns (in-born and out-born) are admitted to the NICU in one year.

411 newborns at $\geq$28 weeks gestation born between January 1995 and December 1996 were recruited. The total number of binary-items available for analysis largely determines the sample size for tool construction. A sample size of approximately 400 subjects for a 66-item index, i.e. six subjects per item, could provide a stability of approximately 80% to parameter estimation in the proposed outcome measure, NMI [91].

3.2.2 Study Design and Study Sample

Since the intent was to employ the IRT model for item analysis and scaling of items on the morbidity continuum, the sampling strategy attempted to achieve a rectangular distribution of morbidity scores in the recruited population. One of the objectives of the study was to *scale* items in the mild to moderately severe range of morbidity, the range of most interest to recent obstetrical interventions. In order to achieve finer discrimination of morbidity in the mild to moderate range, there was a slight over-sampling of the newborn population from this range. Strategies of over-sampling in areas of a continuum,

where discrimination is most required are typical in tool development [68].

The NICU at the Royal Alexandra Hospital admits both newborns delivered at the hospital and transfer admissions from other hospitals in Alberta. For admitting newborns to the NICU at this hospital, there are three sites, A, B, and C. New admissions are admitted to site A or B, depending on the impression of degree of morbidity. Newborns with moderate to severe degree of morbidity are admitted to site A, and newborns with very mild to mild and moderate degree of morbidity are admitted to site B. Site C is a step-down nursery. The newborns initially admitted to the sites A and B are transferred to site C after the acute illness is over and recovery begins.

We recruited 180 newborns representing moderate to moderately-severe degree of morbidity from the site A, 214 newborns representing very-mild to moderate degree of morbidity from the site B, and the remaining 17 newborns representing nil to very mild degree of morbidity from the regular nursery. Recruitment was done by the *consecutive sampling* method [91]. This strategy of recruitment was not only a practical form of recruitment but also the best facsimile of random sampling of newborn morbidity in the context of our study.

3.2.3 Inclusion Criteria

We included newborns at ≥28 completed weeks of gestation with complete maternal delivery and newborn record. There were no exclusion criteria. Newborns with

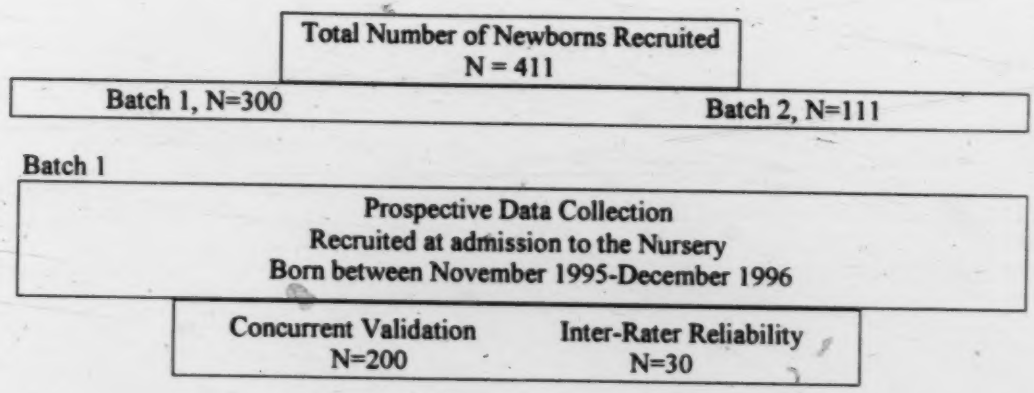
congenital anomalies and metabolic disturbances at birth were also included with an objective to account for morbidity at birth, due to these factors, in the outcome measure.

3.2.4 Data Collection Strategies

Starting November 1995, data were collected over a period of fourteen months. Two batches of samples were obtained between November 1995 and December 1996. The first batch consisted of 300 newborn subjects meeting the inclusion criteria and recruited at the time of admission to the nursery. The first 200 subjects from this batch were employed to obtain data on concurrent validation using a 'visual analogue scale' (Appendix 3). This validation included an overall assessment of morbidity in a newborn marked on a visual analogue scale by the attending physician. We received 165 sheets completed by the attending physicians, which is a compliance of 82.5% on this procedure. The remaining 100 subjects from this batch were proposed for testing reliability on data extraction between the two data collectors, the researcher and the nurse research assistant. A 100% agreement on data collection was obtained on the first 30 charts marked by the two data collectors, blinded from each other. This high degree of agreement was obtained because items in the data collection form are mostly objective, precise, easily extractable and non-ambiguous. It was decided to suspend this testing after obtaining agreement on the first 30 charts of the second batch.

For the remaining 111 newborns in the second batch, data collection was accomplished by chart review of newborns admitted between January 1995 and October 1995, at the Department of Health Records. Data collection by chart review was found to be cost-

efficient without compromising the accuracy, precision and completeness of the collected information. Both data collectors conducted data extraction. This strategy of data collection was employed to advocate that a variety of formats can be employed for data collection on items of this tool. The adjoining flow-chart explains the data-collection strategies employed:



Prior to commencing recruitment, a letter describing the objectives of the study was mailed to all the attending neonatologists and pediatricians (n=38) with admitting privileges at the RAH. We requested their participation in the research project. Their participation consisted of marking an assessment of the overall morbidity on the visual analogue scale rated from 0 to 5, attached to the charts of the recruited newborns. A copy of the analogue scale and the letter mailed to the attending physicians are enclosed as Appendix 3 and Appendix 4, respectively. Following this formal letter of request, we

arranged a meeting with the attending physicians to explain the nature of assessments required and their contribution toward concurrent validation, as a part of external validation. Twenty-five pediatricians and neonatologists with admitting privileges at the RAH attended the meeting, where the purpose of the proposed outcome measure, its contents and the proposed validation procedure were further clarified.

3.2.5 Recruitment of Subjects

A blanket collective consent in the form of ethical acceptability by the Ethics Review Committee, Faculty of Medicine at the University of Alberta was deemed sufficient for recruiting subjects for data collection, since the data collection involved extraction of charted information on the newborns admitted to the nursery. Parents of the recruited newborns were not contacted individually for participation in the study. A copy of the ethical acceptance by the University of Alberta Ethical Review Committee is appended as Appendix 5.

Two people, the researcher and a nursing research assistant, carried out recruitment. They identified infants who met the degree of morbidity criteria, from the *delivery logbook* at the NICU. One of the considerations at the time of recruitment was placement of the newborn in the NICU, i.e. A or B-side, since the placement of the newborn was related to the severity of illness. Upon recruitment, the newborn's file was flagged with a study-label and a sheet of Appendix 3 for the assessment of concurrent validation by the attending physician was attached.

3.2.6 Data Collection

For the structural validation and scaling phase of the study, we collected data on the sixty-six binary items of newborn morbidity. Data were obtained from the newborn's admission sheet, physician's notes, newborn's flow sheet and maternal case sheet. Guidelines for data collectors for extraction of data on the newborn and maternal items are provided in Appendix 6. Appendix 7 contains the form used for recording data on the newborn morbidity items. These newborn data were recorded at the end of day 1, 3, 7 and/or on the day of discharge, whichever was earlier. In the event of a recruited newborn leaving the NICU within 24 hours of birth (e.g., after a few hours of observation at the NICU 'B' side), the newborn was counted as a regular nursery admission.

3.3 External Validation

External validation consisted of correlation between the score on the morbidity index and relevant events that were external to the index. There are two types of external validation events: *cross-sectional and longitudinal*. The cross-sectional events include events occurring in the same time frame as the initial assessments or scoring on the index. Whereas, the longitudinal events occur in the time frame following the initial assessments or scoring. Focused assessments across discrete time frames are more desirable in the case of newly proposed outcome measures.

3.3.1 Events for Cross-sectional External Validation

(i) Concurrent Assessment by Physicians

A sheet of Appendix 3 was attached to the charts of recruited newborns and the charts were flagged. The assessments of overall morbidity in the newborns were marked by the attending physicians on day 1, 3, 7 and/or the day of discharge on the '0-5' visual analogue scale. The analogue scale encompassed the whole range of morbidity continuum, from 'no morbidity to severe morbidity'. We determined the correlation between the average of the physicians' assessments on day 1, 3 and 7 and/or the day of discharge, and the morbidity index.

(ii) Scores on the SNAP and the NTISS

From January 1996 to March 1997, the NICU at the RAH participated in a collaborative project of data collection and validation of the Score for Neonatal Acute Physiology

(SNAP) and the Neonatal Therapeutic Intensity Subsystem Score (NTISS) indices. Both these indices are reflections of morbidity and resource utilization in newborns admitted to the NICU. The former is based on derangement of acute physiology in the first 24 hours, 3 and 14 days, while the latter is a therapy-based index assessed for the entire length of stay in the NICU.

The scores on the SNAP and the NTISS were available for a subset of newborns recruited into our study. We determined the correlation of these scores with the morbidity index on the NMI. We expect to find mid-range correlations of these scores with the NMI.

(iii) Gestational age and (iv) Birth weight

Gestational age and Birth weight were recorded from the delivery record sheet. At the first antenatal visit of a gravida, gestational age is calculated from the certain date of the last menstrual period; when certain dates are not known, an ultrasound examination before 20 weeks of gestation is performed to estimate the expected date of confinement. A newborn is weighed at birth in the case room and the birth weight is recorded to the nearest five gram on the delivery record.

We determined the correlation between the gestational age and the birth weight with the morbidity index. A newborn's gestational age or birth weight is expected to account for only a part of the variance in the morbidity index. The remaining would likely be explained by morbidity *per se*, which is independent of the gestational age and the birth weight.

(v) Maternal Events

Information on maternal events was obtained from the pre-natal and delivery record sheets, and was recorded as shown in Appendix 8. We determined the correlation of the morbidity index with the following maternal conditions, as diagnosed and noted by the attending obstetricians:

(a) Diagnosis of fetal distress

(b) Pre-gestational diabetes mellitus

(c) Gestational diabetes mellitus: i. controlled by diet; ii. controlled by diet + insulin

(d) Preeclampsia

(e) Hemolysis, Elevated Liver Enzyme and Low Platelet (HELLP) syndrome

(f) Maternal Obesity Pre-pregnancy weight >90 kg

(g) Maternal Chronic Diseases i. renal; ii. connective tissue; iii. epilepsy

(h) Maternal Infections: i. chorioamnionitis; ii. Group B Streptococcus 'positivity' (clinical diagnosis of GBS)

(i) Breech Delivery

(j) Antepartum Hemorrhage

(k) Oligohydramnios/Polyhydramnios

Guidelines for data collection on maternal disease conditions are provided in Appendix 6. Other factors influencing newborn morbidity may include maternal socioeconomic status, drug abuse and lifestyle. Due to a complex nature of the definitions of these factors, a considered decision was made to explore these issues separately in future studies.

3.3.2 Events for Longitudinal External Validation

(i) Length of Stay in the Hospital

Length of Stay (LOS) was examined as the total LOS, LOS at the A+B side of NICU and LOS at the C side of NICU (step-down nursery). This was calculated by finding a difference between the date of admission and the date of discharge or the date of transfer to the C-side, whichever the case. We determined the correlation of the morbidity index separately with all the three lengths of stay.

(ii) Neonatal Mortality at 28 days

Correlation between the morbidity index and mortality by 28 days was determined. Information regarding mortality by 28 completed days of life was obtained from the Hospital Health Records. For that reason, only the infants who died before discharge from the hospital were included in this analysis. The investigators assumed that the newborns too morbid during the early neonatal period (i.e. 7 days) not to (possibly) survive until the late neonatal period (i.e. 28 days) would not be discharged from the hospital. Similarly, infants dying after discharge from the hospital (with in 28 days of life) due to conditions other than morbidity at birth, were not suitable candidates for the correlation analysis for external validation of the NMI.

(iii) Postpartum Nursing Utilization

Information on the postpartum nursing utilization was obtained from the records of the Healthy Beginnings Public Health Program of the Edmonton Board of Health.

Postpartum units in Edmonton area are divided into four different zones. A postpartum unit is assigned by the zone of residence of the mother. The hospital authorities inform the postpartum unit regarding an infant's date of birth and date of discharge from the hospital. Following discharge, a Public Health Nurse (PHN) from the unit makes contact with the family with in 24 hours by telephone. The first contact may or may not be followed by subsequent contacts or home-visits depending on the condition of the mother and the newborn. Contacts by the PHN are made to ensure social, mental and physical well being of the mother and the infant. The nature, purpose and outcome of the contact are noted by the PHN on the Case Contact Client Data (CCCD) form.

We defined *utilization* as the *number of contacts* and the duration of *time spent* due to infants' causes by the PHN. A '*contact*' signified a telephone call or home-visit made to the family in the first six weeks of newborn's life. For the purpose of analysis, each home-visit was weighted twice as much as a telephone call. For telephone contacts, the '*time spent*' was computed as the total time spent during the calls. The CCCD form used by the PHN for maintaining the records of contacts and follow-up visits with the family is appended as Appendix 9.

The nurse research assistant was responsible for extracting the charted information from the available centralized records at the Downtown office of the Healthy Beginnings Postpartum Program. The correlations between the morbidity index and *the number of contacts* and the *time spent* were computed.

[Note: Prior to starting data collection for our study, the postpartum program coordinator

at the Edmonton Board of Health informed us that the data on the CCCD forms maintained by PHNs would be entered into a centralized database on computer. During the course of our study, it was realized that the records of the postpartum program could not be centralized or computerized due to administrative reasons, and also the records at the central office were incomplete. However, we were successful in obtaining data on 87 of the 411 recruited newborns. Later, a meeting was held with PHNs from two different zones in Edmonton. The PHNs explained that a number of items in the CCHD form were *subjective measures* (instead of being *hard* and *objective* measures). For these reasons, the researcher proposes in future studies to explore some other avenues of calculating postpartum utilization during infancy.]

Chapter 4

Data Analysis and Results

The data were entered using SPSS 6.1 (SPSS Inc. 1995), and checked using the "valid-entry specification" feature. Two people completed data entry. Random checks were conducted during data collection to ensure comparability of the information between the hard copy and the computer version. Other checks using frequency distribution command were conducted to ensure correct data entry.

Descriptive analyses and parts of the inferential analyses were conducted using SPSS/PC+ Base and Advanced Statistics, version 6.1 (SPSS Inc., 1995). DIMTEST software [80] was used for dimensionality testing. 'Item Analyses and Calibration' of the binary (logisitic) data were conducted using the BILOG-W program [83].

4.1 Descriptive Statistics

4.1.1 Subjects

The recruited sample consisted of newborns at $\geq$ 28 weeks gestation, ranging in morbidity from nil to moderately severe, born between January 1995 and December 1996 and admitted to a nursery at the Royal Alexandra Hospital. At the time of data collection, an estimated degree of morbidity in each newborn subject was ascertained from the sum of the number of items scored. All through the data collection phase, a 'frequency histogram' of the number of items scored (similar to that shown in Figure 4.1) was used

as a guide for determining the number of subjects for recruitment in each of the three broad categories of mild, moderate and severe morbidity. The final sample for the structural validation phase consisted of 411 newborns, 231 males and 180 females. The newborn characteristics are summarized in Table 4-1. Percentile values of newborns across gestational age (GA) and birth weight strata are shown in Table 4-2.

Table 4-1.

Newborns: Birth Weight and Gestational Age

Characteristics	N	Minimum	Maximum	Mean	Std. Deviation
Birth Weight (grams)	411	685	5085	2550.97	941.56
Gestational Age (weeks)	411	28.0	42.4	35.78	3.78

Table 4-2.

Newborns: Percentile Values

Percentile Value	25^{th}	50^{th}	75^{th}
Birth Weight (grams)	1750	2500	3240
Gestational Age (weeks)	32	35.7	39.3

The maternal and delivery characteristics are summarized in Table 4-3 and Table 4-4.

Table 4-3.

Maternal Characteristics

	N	Minimum	Maximum	Mean	Std. Deviation
Maternal Age	411	16	42	28.38	5.96

Table 4-4.

Frequency Distribution of Maternal and Delivery Characteristics

Maternal Characteristic	Number of Subjects	Percentage
HELLP Syndrome	6	1.5 %
Pre-gestational diabetes mellitus	10	2.4 %
Gestational diabetes mellitus (controlled with diet only)	15	3.6 %
Gestational diabetes mellitus (controlled with diet and insulin)	15	3.6 %
Oligohydramnios	24	5.8 %
Antepartum hemorrhage	26	6.3 %
Maternal obesity (Pre-pregnancy weight >90 kg)	43	10.5 %
Fetal distress	65	15.4 %
Preeclampsia	72	17.5 %

4.1.2 Summary of Modifications to the Item-List

Modifications to the morbidity item-list were made at various stages during the substantive and structural validation phases of the study. The purpose of the modifications during the substantive phase was to warrant clarity, accuracy and representativeness of the selected items. Simultaneously, checks were made to remove redundancy in the collected information due to overlapping items. Modifications during the structural validation phase were dictated by the behaviour of included items as it relates to the structure of newborn morbidity. The goodness-of-fit criteria were examined for interpretations of suitability of items in the final scale. Table 4-5 summarizes the modifications made to the provisional item-list during the validation phases.

Table 4-5.

Summary of Modifications of Items in the Newborn Morbidity Index

Number of items available for substantive validation = 60

Number of items available for preliminary structural validation = 66

Number of items available for final stage 1 of structural validation = 58

Number of items available for final stage 2 of structural validation = 50

Number of items available for the scaling study = 50

Attribute and Serial No.	Modification	Binary Item Number
(For item address: Refer to Table 3-3, Columns Serial Number and Binary Item Number)		
(i) After the Preliminary-Testing Phase		
1. Respiratory rate/minute (H.a, H.b)	**Modify**: R.R./minute <30 or >60 in 24 hr. to < 30 or > 60/minute in the first 3 hr. and < 30 or > 60/minute in 3-24 hr.	(17 and 18)
2. Flaccidity (K.a)	**Add**: Flaccidity	(25)
3. Altered Colour (I.a, I.b)	**Alter**: Peripheral cyanosis/Central Cyanosis to Ruddy/Peripheral cyanosis and Dusky/Central cyanosis	(20 and 21)
4. Level of Consciousness (O.a)	**Alter**: Hyperalert to Hyperalert/Hypertonic /Hyperirritable	(35)

(ii) After the Feasibility Phase

1. Intra-Ventricular Hemorrhage (X.a, X.b)	**Add**: Intraventricular Hemorrhage	(65 and 66)

(iii) After the Data Collection Phase

1. Apgar Score of 0 at 10 minutes (E.c)	**Delete**	(12)
2. Facial/Peripheral Nerve Injury (T.d)	**Delete:** Injury present during the first 24 hours only	(52)
3. Spinal Cord Injury (T.i)	**Delete**	(57)

(iv) After the Preliminary-Item-Analysis using BILOG

1. Long bone/Clavicle fracture and Skull fracture (T.b, T.c)	**Combine**: Long Bone/Clavicle/Skull fracture	(50 and 51)
2. Facial or Peripheral Nerve Injury (T.e, T.f)	**Alter:** Facial/Peripheral Nerve Injury – yes/no	(52, 53 and 54)
3. Subdural Bleed and Intracerebral Bleed (T.g, T.h)	**Combine:** Subdural/Intracerebral Bleed	(55 and 56)

(v) After the Final Item Analysis Stage 1 using BILOG

1. Cord Blood pH (A.a, A.b)	**Combine**: pH < 7.1 and pH < 7 into pH < 7.1	(1 and 2)
2. Hypoglycemia (U.a, U.b)	**Combine:** Blood glucose <2.2 mmol/L and Blood glucose <1.7 mmol/L into Blood glucose <2.2 mmol/L	(58 and 59)

(vi) After the Final Item Analysis Stage 2 using BILOG

1. Resuscitation at Birth (B.a)	**Delete:** Bag & Mask at birth	(3)
2. Meconium above cords (C.a)	**Delete**	(5)
3. Respiratory Rate (R.R.) (H.a)	**Delete**: R.R./minute <30 or >60 in first 3 hr.	(17)
4. Altered Colour (I.a)	**Delete**: Ruddy/Peripheral Cyanosis	(20)
5. Hypotonia (J.a)	**Delete**: Identified at <1 hour of age	(23)
6. Level of Consciousness (O.a)	**Delete**: Hyperalert/Hypertonic/Hyperirritable	(35)
7. Cephalhematoma (T.a)	**Delete**	(49)

After data collection, the items *Apgar score of 0 at 10 minutes* and *Spinal cord injury* were deleted since none of the recruited newborns scored on these items. The categories of the item *Facial/Peripheral nerve injury* (*within 24 hours only* vs. *nerve injury - improving vs. not improving at discharge*) were modified to produce a single binary item of *Facial/Peripheral injury vs. no such injury*. The primary reason for this modification was that there were none to few subjects in the individual categories of the attribute *Facial/Peripheral nerve injury - present during the first 24 hours only* (n, number of subjects=0) *vs. injury present and improving at discharge* (n=3) *vs. injury present and not improving at discharge* (n=1). Similarly, due to a small number of subjects in the individual categories, the items *Long Bone/Clavicle Fracture* (n=4) and *Skull Fracture* (n=1), and the items *Subdural Bleed* (n=4) and *Intracerebral Bleed* (n=2) were clumped together. This modification was based on the fact that morbidity due to individual items in each of the two sets is quite similar to each other. Accordingly, four items were clumped to produce two binary items. Based on the results of the two-parameter estimation and calibration during item analysis of the 58-item index, seven items were deleted (refer Table 4-6, column 58-item BIN for the item names). The explanations for these deletions are summarized in Table 4-9, and the plots of their item response functions are compiled in Appendix 11. As a result of these modifications, the final morbidity item-list consisted of 50 binary items.

4.1.3 Descriptive Statistics: Items

The raw score frequencies for all the 50 items are given in Table 4-6. A striking feature of these scores is the frequency of 'assumed continuum of morbidity' for certain items. A few examples of this feature are: Item 30 and Item 31 are assumed to have been scored if a newborn scores on Item 32, i.e. poor sucking beyond seven days of life would be preceded by poor sucking at <24 hours and <7 days after birth. Item 24 and Item 25 are assumed to have been scored if a newborn scores on Item 26, i.e. a newborn must have a single seizure followed by multiple seizures before being considered for therapy with more than two drugs. The column '58 Item No.' in the Table 4-6 represents the binary item numbers given to the morbidity items for the 58-item morbidity index and the column '50 Item No.' represents the binary item numbers given to the morbidity items for the final 50-item morbidity index. The latter binary item numbers will be used in all subsequent tables to depict the results of item parameter estimations by both one and two parameter models. The columns 'Record' and 'Score' represent the number of subjects for which recorded data were available for an item and the number of subjects scoring on that item, respectively.

In the sample population, the minimum number of items scored by any newborn was 0, and the maximum number was 28. Table 4-7 contains the frequency distribution of the number of items scored. Figure 4-1 is a graphical representation of the same distribution.

Table 4-6.

Raw Score Frequency
BIN for Index: 58-Item and 50-Item

SNo.	Morbidity Item	58-Item BIN	50-Item BIN	Record	Score
A	**Cord Blood pH**				
a	value <7.1	1	1	211	28
B	**Resuscitation at birth**				
a	Bag & mask	2	delete		
b	Intubation	3	2	411	89
C	**Meconium**				
a	Meconium only above cords	4	delete		
b	Meconium above and below cords	5	3	411	22
D	**Apgar Score (5 minute)**				
a	score <7	6	4	410	95
b	score<4	7	5	410	15
c	score<1	8	6	410	2
E	**Apgar Score (10 minute)**				
a	score <7	9	7	235	24
b	score<4	10	8	235	3
F	**Heart Rate/minute**				
a	>160/minute	11	9	411	117
b	>200/minute	12	10	411	4
c	<100/minute	13	11	411	20
G	**Systolic BP (mean, mm of Hg)**				
	28-32 weeks 32-42 weeks				
a	<30 <40	14	12	411	60
H	**Respiratory Rate**				
a	R.R./minute <30 or >60 in the first 3 hr.	15	delete		
b	R.R./minute >100 between 3-24 hr.	16	13	411	48
c	R.R./minute <30 or >60 between 3-24 hr.	17	14	411	72
I	**Colour**				
a	Altered Colour OR Ruddy	18	delete		
b	Central Cyanosis OR Dusky	19	15	411	39

Table 4-6.

Raw Score Frequency
BIN for Index: 58-Item and 50-Item

SNo.	Morbidity Item	58-Item BIN	50-Item BIN	Record	Score
J	**Hypotonia**				
a	Identified at <1 hr. of age	20	**delete**		
b	Persisting (or Identified) at >1 hr. of age	21	16	411	82
c	Persisting at >120 hr. of age	22	17	411	21
K	**Flaccidity**				
a	Present between 1-120 hr.	23	18	411	24
L	**Apnea**				
a	Apneic spells (by apnea monitor)	24	19	411	237
b	Apneic spells and need for oxygen	25	20	411	139
c	Apneic spells and need for resuscitation	26	21	411	50
M	**Bleeding disorder**				
a	Thrombocytopenia OR Bleeding Skin/Lungs/GIT	27	22	411	34
b	Need for transfusion due to disorder M.*a*	28	23	411	9
N	**Seizures**				
a	Tremors OR Non-Recurring Single Seizure	29	24	411	56
b	>1 seizure	30	25	411	22
c	If >2 drugs used for treatment	31	26	411	4
O	**Level of consciousness**				
a	Hyperalert/Hypertonic/Hyper-irritable	32	**delete**		
b	Drowsy OR Lethargic	33	27	411	57
c	Stupor OR Obtundation OR Coma	34	28	411	3
P	**Cardio-Pulmonary Resuscitation**				
a	Any time before discharge	35	29	411	15
Q	**Oral Feeding Difficulties**				
a	Poor sucking at <24 hr.	36	30	411	300
b	Poor sucking between 24 hr. - 7 days	37	31	411	222
c	Poor sucking beyond day 7	38	32	411	118
d	Persistent vomiting	39	33	411	59

Table 4-6.

Raw Score Frequency
BIN for Index: 58-Item and 50-Item

SNo.	Morbidity Item	58-Item BIN	50-Item BIN	Record	Score
R	**Respiratory Status**				
a	Assisted ventilation in the first 24 hr.	40	34	411	188
b	Assisted ventilation after 24 hr.	41	35	411	130
c	Mechanical ventilation in the first 24 hr.	42	36	411	79
d	Mechanical ventilation between 24 hr. - 7 days	43	37	411	63
e	Mechanical ventilation beyond day 7	44	38	411	16
S	**Urine Output**				
a	Abnormal (low)	45	39	411	48
T	**Trauma at birth**				
a	Cephalhematoma	46	delete		
b	Bone fracture including long bone OR clavicle OR	47	40	411	6
	Facial OR Peripheral Nerve Injury				
c	Present within 24 hr. of birth	48	41	411	4
d	Present beyond 24 hr. of birth	49	combined		
e	Subdural OR Intracerebral Hematoma	50	42	411	5
U	**Hypoglycemia**				
a	Blood Glucose <2.2 mmol/L	51	43	411	27
V	Hyperbilirubinemia (peak level-micromol/L)				
a	Serum bilirubin >170	52	44	411	231
b	Serum bilirubin >250 OR Phototherapy	53	45	411	195
c	Serum bilirubin >340 OR Exchange transfusion	54	46	411	5
W	**Bacterial Culture**				
a	Blood Positive	55	47	411	15
b	CSF Positive	56	48	411	3
X	**Intra-ventricular Hemorrhage**				
a	Grade 1 and 2	57	49	411	26
b	Grade 3 and 4	58	50	411	4

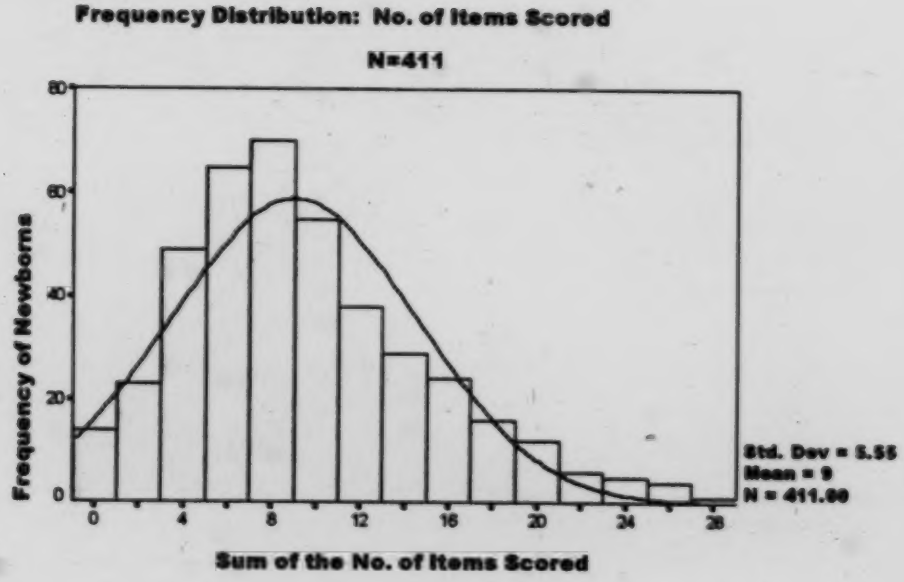
SNo. = Serial Number; BIN = Binary Item Number; Record = Number of recorded readings on the item; Score = Number of newborns scoring on that item.

Table 4-7.

Frequency Distribution of the Number of Items Scored by Newborns

Mean	Median	Mode	Minimum	Maximum	SD
9	8	7	0	28	5.5
Percentile Value		25th	50th	75th	
		5	8	12	

Figure 4-1.



4.2 Inferential Statistics

4.2.1 Dimensionality of the Data Set

Before undertaking item parameter estimation, dimensionality of the latent trait in the data set was determined using DIMTEST. The null hypothesis for this test is that newborn morbidity in the data set is unidimensional. Items for the initial subset (assessment 1 subset, AT1 discussed on page 23) for analysis were selected by both content-based knowledge of the trait of morbidity (Set A) and exploratory data analysis (Set B) method. Table 4-8 contains the results pertaining to the determination of dimensionality performed using data from 411 newborns. Various combinations of the 'number of items in a set' (subset AT1) and the 'minimum number of subjects in each cell for computation' were examined.

Table 4-8.

DIMTEST Statistics for Determination of Dimensionality

Set	Min. Cell Number*	% of Population	T-Statistic	p-value
A. User Supplied - AT1 Set (Content-Based Set)				
1	5	92	0.72	0.24
2	10	91	0.72	0.24
3	15	85	0.28	0.39
4	20	82	0.35	0.36
B. Exploratory Data Analysis - AT1 Set				
1	5	90	-2.8	0.99
2	10	86	-2.9	0.99
3	15	83	1.83	0.03
4	20	80	0.28	0.39

*: Minimum number of subjects in a cell for computation

A p-value of greater than 0.05 signifies that the null hypothesis is tenable. The results in Table 4-8 illustrate that a p-value of 0.03 (i.e. $p<0.05$ for the 'B-3' Set) was obtained only once in the eight iterations reported. From the results of this analysis, it was clear that a single dimension provided an excellent fit to these data, as evidenced by a low T-value statistic and a $p>0.05$. The p-value of >0.05 was obtained for approximately 90% of the total of 25 iterations that were performed. These results were consistent with the hypothesis that there was a single construct underlying the data; and that was probably morbidity. Based on the results of this analysis, we elected to perform parameter estimation and scale modeling within the framework of IRT.

4.2.2 Structural validation and Scaling of the Items

Both one and two-parameter IRT approaches using the BILOG program were employed to determine the parameters and sequence of items on the morbidity continuum. Both models seemed to provide a similar pattern of sequencing of the items. However, greater information was furnished by the latter model, in terms of the number of parameters for individual items and the total test information function.

The goodness-of-fit analyses were performed using both one and two parameter model techniques on the 58-item index. Table 4-9 contains the results of this analysis. Results using the two-parameter model showed that eight items had a relatively poorer fit on the morbidity continuum. The criteria employed for the goodness of fit were the chi-square statistic, discrimination statistic, plots of item response function curves (Appendix 10),

extreme unwarranted location of the item on the 'difficulty' scale, and content review in light of the results of the parameter estimation. Seven of these eight items were deleted. The item *assisted ventilation in first 24 hours* (Item 34) was retained in spite of a poor fit, because it was deemed to be at a suitable location on the morbidity scale, and did not possess a negative information function. The items that fit poorly on the morbidity continuum had a high chi-square and dispersion statistic, a low discrimination and item information function. Each of the seven deleted items, such as meconium above cords, bag and mask at birth, respiratory rate <30 OR 60-100/minute in the first 3 hours, ruddy/peripheral cyanosis, hypotonia in the first one hour after birth, hyperalert/hypertonic/hyperirritable state and cephalhematoma, signify a very mild morbidity. Examination of the item response curves of these items confirmed a 'poor fit' and a negligible 'information function'. The information curve of all these items lacked slope, the 'a' parameter, and the probability of scoring on these items did not vary along the continuum of morbidity. Appendix 11 contains the item response curves of the seven deleted items. Most importantly, the content review of the deleted items revealed that those items were factually not true images of morbidity. Descriptively, those items probably reflect opinions of slight aberrations of physiology at birth or that of suspicion or anticipation of morbidity by the treating physician. The inclusion of those items contributed to more error than information to the scale values, along the morbidity continuum. Further, the modified item *Facial/Peripheral Nerve Injury* now had two categories, i.e. injury vs. no injury. Accordingly, a morbidity index consisting of 50 binary items was obtained for the final item-analysis for computing the sequence and scale values of the included items.

Table 4-9.

Goodness-of-Fit Criteria
Morbidity Items in Ascending Order by the Difficulty Parameter
58-Item Index

Two Parameter IRT Model				One Parameter IRT Model		
BIN	Discrim.	Diffi.	Chi Sq.	BIN	Diffi.	Chi Sq.
36	1.14	-0.80	10	36	-1.29	13.4
52	0.39	-0.39	22.2	24	-0.41	12.6
24	0.87	-0.25	8.5	52	-0.34	26.8
37	1.46	-0.13	5.3	37	-0.32	48.7
53	0.41	0.18	21.3	53	0.10	24.9
25	1.49	0.53	6.7	40	0.18	97.4
40	0.19	0.57	65.1	15	0.51	82.7
38	1.15	0.74	12.9	2	0.62	95
41	0.76	0.80	26.2	25	0.79	41.1
11	0.60	1.10	2.9	41	0.91	24.7
42	1.27	1.11	2.8	38	1.04	32.6
3	0.99	1.12	8	11	1.09	7.7
21	1.12	1.13	8.8	6	1.43	7.8
43	1.45	1.23	4.5	3	1.53	4.4
15	0.19	1.40	37.5	21	1.65	15.5
45	1.47	1.41	1.5	42	1.70	18.9
14	1.00	1.48	7.9	4	1.79	50.3
33	1.08	1.49	0.6	17	1.83	13.9
26	1.19	1.50	3.4	18	1.89	26.3
9	1.64	1.52	2.3	43	2.02	17.6
6	0.51	1.59	14.1	14	2.09	13.3
29	0.92	1.61	2.2	39	2.11	14.7
17	0.67	1.67	16.8	29	2.18	7.2
30	2.05	1.70	1.8	33	2.18	17.5
22	1.82	1.78	1.4	20	2.20	42.3
35	2.28	1.83	2.3	26	2.32	14.6
39	0.68	1.88	11.8	16	2.38	3.2
23	1.25	1.96	3.7	45	2.38	18.4
27	0.94	2.01	2.3	1	2.39	6.6
19	0.81	2.07	17.9	9	2.60	11.6
44	1.35	2.12	1.9	19	2.64	5.2
57	0.96	2.19	1.3	27	2.80	8.9

Table 4-9.

Goodness-of-Fit Criteria
Morbidity Items in Ascending Order by the Difficulty Parameter
58-Item Index

Two Parameter IRT Model				One Parameter IRT Model		
BIN	Discrim.	Diffi.	Chi Sq.	BIN	Diffi.	Chi Sq.
31	2.53	2.26	0.3	51	3.08	1.4
7	1.14	2.35	1.4	57	3.12	10.9
16	0.59	2.35	8.5	23	3.21	16
2	0.14	2.37	35.1	5	3.31	10
10	1.85	2.43	0.2	30	3.31	15.2
58	1.74	2.49	0.4	22	3.37	10.8
1	0.54	2.52	9.6	13	3.42	7
34	1.91	2.53	0.2	46	3.54	25.1
28	1.23	2.53	0.8	32	3.67	14.8
55	0.94	2.62	1.5	44	3.67	7.9
8	1.57	2.88	0.1	7	3.74	3.6
56	1.03	3.47	0	55	3.74	1.1
51	0.48	3.55	2.2	35	3.74	10.7
13	0.54	3.58	2.4	28	4.30	0.6
50	0.81	3.73	0	47	4.73	0
18	0.26	3.79	13.5	50	4.93	0
47	0.74	3.84	0	54	4.93	0
12	0.72	4.28	0	10	4.94	0
49	0.95	4.37	0	12	5.16	0
5	0.40	4.54	10.9	31	5.16	0
4	0.19	4.85	18.4	48	5.16	0
20	0.23	5.00	7.2	58	5.16	0
48	0.53	5.56	0	34	5.46	0
54	0.45	6.11	0	56	5.46	0
32	0.31	6.33	4	8	5.88	0
46	0.24	7.78	14.4	49	6.60	0

BIN = Binary item number for the 58-item index; Discrim.= Discrimination statistic; Diffi.= Difficulty statistic; Chi Sq.= Chi-Square statistic.
Items with poor-fit and relevant parameter statistic showing a poor fit are boxed.

Using both one and two parameter models of the IRT, item analysis of the 50-item morbidity index was performed on the data set of 411 newborns. The test information and the measurement error curves for the two models are illustrated in Figures 4-2 and 4-3. The solid ascending line in the graph represents the 'information function' of the index, and the dotted line corresponds to the standard error of estimate (SEE) along the continuum of morbidity. The plots suggested that the two-parameter model provided a better fit to the data, particularly in the moderate morbidity zone. It is apparent that the SEE is least in the range where the information is maximum. For the one-parameter model, the information curve encompasses a range of morbidity from –1 to 3+. In contrast, the information function for the two-parameter model is higher and sharper in the morbidity range of +1 to 3+. Accordingly, the SEE is much lower in that range for the two-parameter model. A lower SEE results in narrower confidence intervals, and therefore, statistically more powerful results. In contrast to the two-parameter model, the one-parameter model provided more information in the mild morbidity zone, as illustrated in Figures 4-2 and 4-3.

Figure 4-2.

One Parameter Model of IRT: BILOG

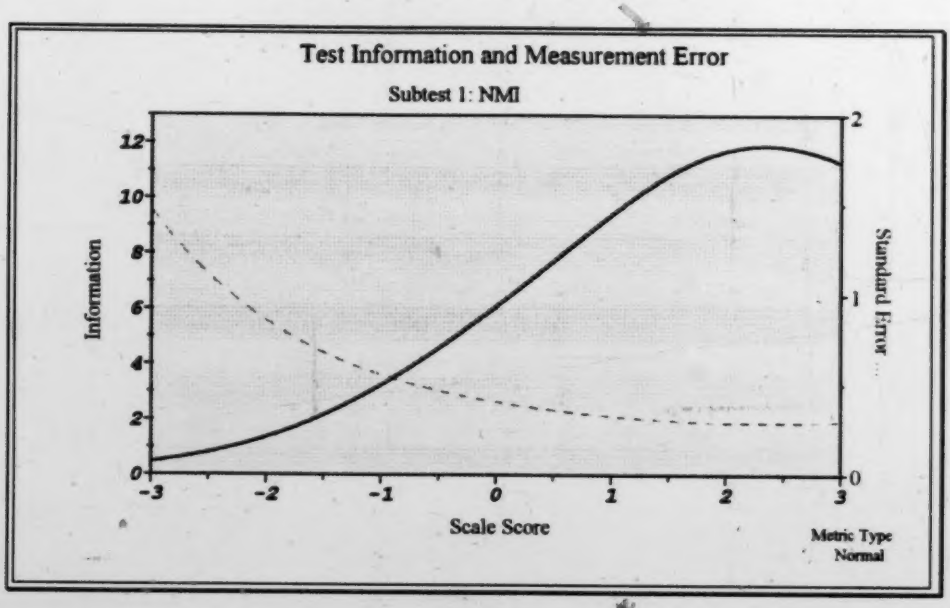


Figure 4-3.

Two Parameter Model of IRT: BILOG

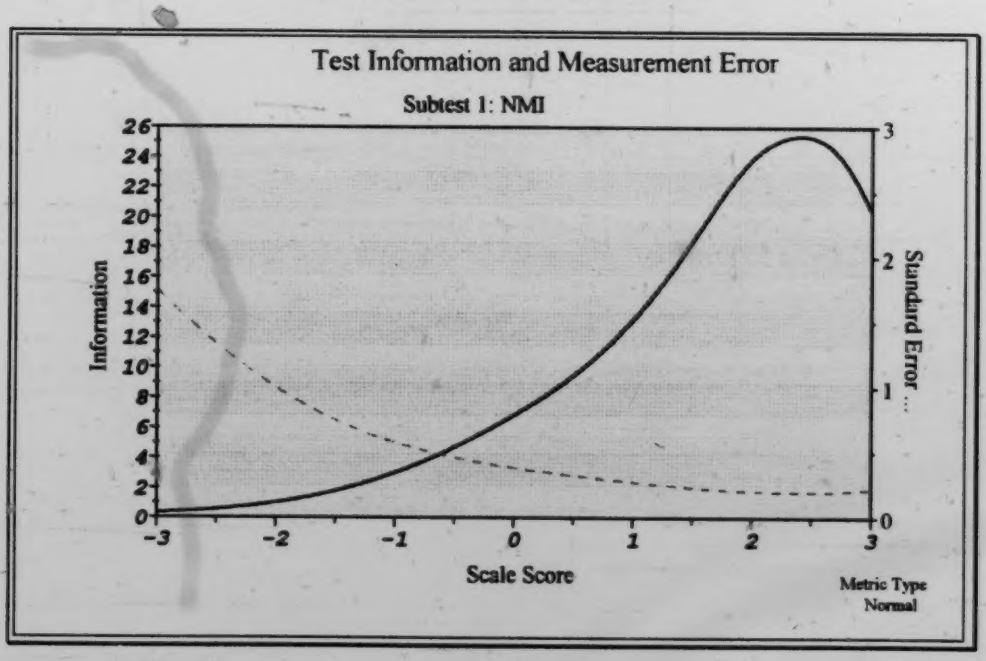


Table 4-10 contains the performance characteristics of the morbidity index including average information function, reliability coefficient and standard error of estimate (SEE) by both one and two parameter models.

Table 4-10.

Performance Characteristics of Models

Model	One-Parameter Model	Two-Parameter
Average Information Statistic	2.2	2.9
Average SEE	0.69	0.59
Reliability Index	0.86	0.89

Average information is the information function integrated with respect to a 0, 1 normal distribution of morbidity. The standard error function is the square root of the reciprocal of the information function. The reliability index is a measure of the internal consistency of the included items in a tool. It is computed from the squared error averaged with respect to the population distribution of morbidity. It is equal to the population variance divided by the sum of the population variance and the mean square error. It is equivalent to Cronbach's alpha statistic but it estimates the actual reliability and not its lower bound. It reflects that items contained in the tool are affected in the same or related ways in response to various influences. For instance, a newborn of a diabetic mother with uncontrolled plasma glucose levels may emulate a full blown picture of severe morbidity with features of low apgar score, respiratory distress, acidosis, hypoglycamia, oral feeding difficulties, altered sensorium, hypotonia, clavicle or long bone fracture and

seizures etc., whereas a newborn of a mother with well controlled plasma glucose may have only a few of these features, and/or that too transiently.

The two models were similar to each other in their performance characteristics. However, the average SEE by the one-parameter model was marginally higher, and information and reliability coefficient were marginally lower as compared to the two-parameter model.

The BILOG was also used to plot the item response function and the goodness-of-fit curves for all the 50 included morbidity items. The same is provided in Appendix 10. The solid ascending line in those curves represents the probability of endorsing an item due to morbidity, and the dotted curve represents the information function of the item. The 'a' parameter corresponds to the 'discrimination statistic' and the 'b' parameter noted on the graph depicts the 'difficulty statistic' of the item on the continuum. Item-analysis furnished the item parameter estimates of 'discrimination' and 'difficulty', and the subject parameter estimates of the morbidity statistic (θ) in a newborn.

4.2.2.1 Item Discrimination Parameter

This analysis included estimates of the item discrimination parameter of the 50-item morbidity index obtained using the two-parameter model, and the results are summarized in Table 4-11. The one-parameter model assumes this statistic to be a constant and the model assigned it a value of 0.75. It is theoretically possible for the discrimination estimates to vary from negative to positive infinity. However, in practice they usually take on values between zero and two. There was a wide variability between items on the

discrimination parameter, with relatively poor fitting items having a value less than 0.50 and better fitting items having a sharper discrimination capability at value greater than 1.5. The larger the discrimination parameter, the lower the dispersion. The two-parameter model incorporates discrimination estimates while estimating the difficulty statistic of an item.

In Table 4-11, the twelve items possessing the highest discrimination parameter at value ≥1.30, in the descending order of the statistic were *cardio-pulmonary resuscitation (anytime before discharge)* [Item 29], *multiple seizures controlled by two or more drugs* [26], *multiple seizures* [Item 25], *apgar score <4 at 10 minutes* [Item 8], *apgar score <1 at 5 minutes* [Item 6], *stupor/obtundation/coma* [Item 28], *intraventricular hemorrhage-grade 3 and 4* [Item 50], *apgar score <7 at 10 minutes* [Item 7], *hypotonia between 1-120 hours* [Item 17], *mechanical ventilation between 24 hours-7 days* [Item 37], *low urine output* [Item 39] and *poor sucking between 24 hours-7 days* [Item 31]. [Please refer to the 'Check' column in Appendix 12 for the Binary Item Numbers]. The discrimination statistic is taken into consideration while scaling the items by the two-parameter model of the IRT. This statistic provides distinction between two newborns in the level of morbidity along the morbidity continuum.

Similarly, the items possessing the least discrimination at value ≤0.50, in the ascending order of the statistic were *assisted ventilation in the first 24 hours* [Item 34], *serum bilirubin between 170-250 units* [Item 44], meconium *above and below cords* [Item 3], *serum bilirubin between 251-340 units OR need for phototherapy* [Item 45],

hypoglycemia (plasma glucose <2.2 mmol/L) [Item 43], *serum bilirubin >340 units OR need for transfusion* [Item 46], *apgar score <7 at 5 minutes* [Item 4] *and birth trauma-facial/peripheral nerve injury* [Item 41]. A low statistic may be an indication of a poor fit due to a poor reflection of the morbidity concept by the item or poor comprehension of the item due to ambiguous or unclear wording or inadequate number of subjects (scoring on the item) for IRT to furnish appropriate estimates. Item 34 and Item 44 may be examples of the first type, i.e. poor reflection or hint of morbidity by the item; Item 3 and Item 45 are examples of the second type, i.e. ambiguity. While collecting data on Item 45, it was found that among other indications for phototherapy, admission to the NICU for observation or otherwise increased the possibility of treatment by phototherapy, besides that due to a truly high level of the serum bilirubin. In retrospect, this item could have performed better if stringent criteria were applied in the formation of this item. For example, for the diagnosis of hyperbilirubinemia and susequent phototherapy, the *serum bilirubin level of 171-250 units* was clumped with the *need for phototherapy*. Rather than clumping the two, a proposal to apply different cut-off levels of serum bilirubin for different gestational ages could have served the purpose better. The examples of the third type include Item 46 and Item 41. Because of the small number of subjects in the categories of these items, it is possible that analysis may fail to provide a stable or clinically acceptable location of these items on the morbidity continuum.

Resolution lies in the exclusion of items with low information and/or re-framing of the ambiguous items and/or including larger sample size (up to 15-20 subjects/item) for utilizing the full benefit of the two-parameter model for parameter estimation.

Table 4-11.

Discrimination Parameter
Two Parameter Model
Number of Items=50

BIN	Discrimination	BIN	Discrimination
1	0.51	26	1.97
2	0.89	27	0.93
3	0.36	28	1.59
4	0.47	29	2.03
5	1.00	30	1.04
6	1.62	31	1.31
7	1.52	32	1.04
8	1.71	33	0.63
9	0.54	34	0.16
10	0.64	35	0.66
11	0.50	36	1.09
12	0.86	37	1.31
13	0.55	38	1.20
14	0.60	39	1.31
15	0.76	40	0.70
16	1.00	41	0.49
17	1.37	42	0.72
18	1.08	43	0.44
19	0.79	44	0.35
20	1.27	45	0.37
21	1.05	46	0.45
22	0.84	47	0.84
23	1.08	48	0.94
24	0.84	49	0.76
25	1.77	50	1.58

BIN = Binary Item Number;
5 items with the highest discrimination parameter are aligned to right.
5 items with the lowest discrimination parameter are aligned to left.

4.2.2.2 Item Difficulty Parameter and Scaling of the Items

Table 4-12 contains the results of the application of both one and two-parameter IRT models to determining the sequence and scale values of the 50 items on the continuum of newborn morbidity. The difficulty parameters as estimated by both models are reported. The sequence and the scale values of the items on the difficulty scale determine the amount of estimated morbidity (θ) in a newborn. Please refer to Table 4-14 and 4-15 for item names and location corresponding to the BIN.

Following estimation of the difficulty parameter, it was linearly transformed by rescaling of the morbidity parameter (θ) to a mean of 100 and a standard deviation (SD) of 20. This linear transformation was performed to facilitate interpretation by eliminating negative scale values on the difficulty parameter and hence on the morbidity scale. It can be interpreted from the results in column 1 and column 4 of the Table 4-12 that the sequence of items in the two models was similar for the most part, but was not identical.

Table 4-12.

Morbidity Items in Ascending Order by the Difficulty Parameter
N = 50 Items

Two Parameter Model			One Parameter Model		
BIN	Diffi. Para.	Rescaled Value	BIN	Diffi. Para.	Rescaled Value
30	-0.84	87	30	-1.06	83
44	-0.43	94	19	-0.34	96
19	-0.27	96	44	-0.28	97
31	-0.08	99	31	-0.18	98
45	0.21	104	45	0.10	103
20	0.60	111	34	0.17	104
34	0.67	112	20	0.71	113
32	0.84	115	35	0.81	115
35	0.91	116	32	0.95	118
9	1.23	121	9	0.96	118
2	1.25	122	4	1.25	123
36	1.26	122	2	1.34	124
16	1.27	122	16	1.44	126
37	1.40	124	36	1.48	127
39	1.61	128	14	1.59	128
27	1.69	129	37	1.75	131
12	1.70	129	12	1.81	132
21	1.71	129	33	1.82	132
7	1.72	130	27	1.86	133
4	1.76	130	24	1.88	133
24	1.82	131	21	2.01	135
14	1.90	133	13	2.05	136
25	1.97	134	39	2.05	136
33	2.08	135	1	2.11	137
29	2.08	136	7	2.25	140
17	2.13	136	15	2.27	140
22	2.26	139	22	2.41	142
18	2.26	139	43	2.63	146
15	2.29	139	49	2.67	147
38	2.44	142	18	2.75	148
49	2.47	142	3	2.83	149
13	2.59	144	25	2.83	149
26	2.66	145	17	2.87	150
50	2.70	146	11	2.91	151

Table 4-12.

Morbidity Items in Ascending Order by the Difficulty Parameter
N = 50 Items

Two Parameter Model			One Parameter Model		
BIN	Diffi. Para.	Rescaled Value	BIN	Diffi. Para.	Rescaled Value
5	2.71	146	38	3.12	154
8	2.77	147	5	3.18	155
1	2.78	147	47	3.18	155
23	2.83	148	29	3.18	155
28	2.85	148	23	3.63	163
47	2.91	149	40	3.97	169
6	3.11	153	46	4.13	171
48	3.89	166	42	4.13	171
43	3.93	166	8	4.16	172
11	3.94	167	41	4.32	175
40	4.14	170	10	4.32	175
42	4.25	172	26	4.32	175
10	4.66	179	50	4.32	175
3	5.11	186	28	4.56	179
41	6.00	201	48	4.56	179
46	6.60	211	6	4.89	184

Refer Table 4-6, column 50-item BIN for item names corresponding to the BIN; BIN = Binary Item Number; Diffi. Para. = Difficulty parameter; Rescaled values = Rescaled difficulty parameter on assigning a mean of 100 and a SD of 20 to the morbidity index.

4.2.2.3 Estimate of Morbidity (θ)

Based on the estimates of difficulty and discrimination parameters, the BILOG analysis provided the estimates of morbidity parameter for both one and two parameter models. We rescaled the morbidity parameter to a mean of 100 and a SD of 20 using the BILOG program. Figures 4-4 and 4-5 are the graphical representation of the frequency distribution of thus computed newborn morbidity index scale values, which we have named as *Newborn Morbidity Index*, NMI.

The results of the frequency distribution, percentile values, and 95% confidence interval (CI) around the morbidity parameter (θ) computed from the standard error of estimate for both models are summarized in Table 4-13. As described on page 29-30 in Chapter 2, and illustrated in Figures 4-2 and 4-3, the SEE varies along the morbidity scale, being minimum in the region of maximum test information and vice versa. The 95 % CIs derived from the SEE were computed at four percentile cut points of the morbidity index, and are given in Table 4-13.

The percentile rank scale values of the NMI were obtained at each 20th percentile. The percentile ranks are used for making relative evaluations of an individual's trait. One disadvantage of using only percentile rank scales and ignoring other systems of scoring is that this scale itself is ordinal and its units are unequal on the continuum of the scale.

Figure 4-4.

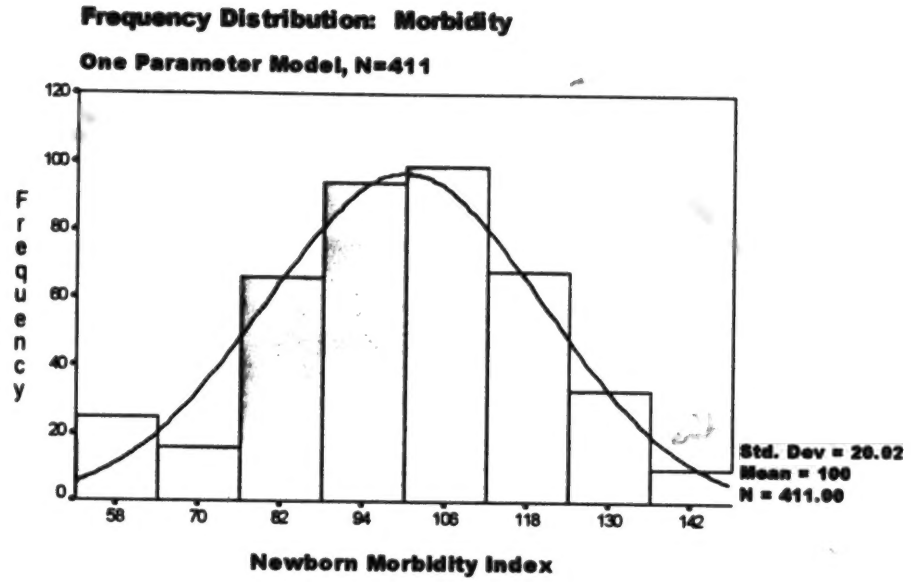


Figure 4-5.

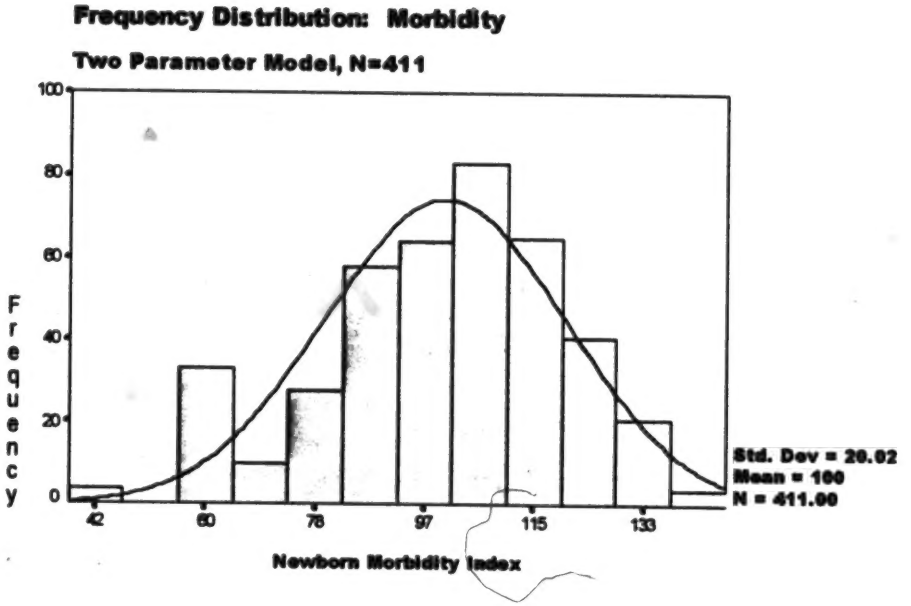


Table 4-13.

Frequency Distribution, Percentile Values and 95% Confidence Interval
Newborn Morbidity Index

Number of Subjects = 411, Number of Items = 50

	One-Parameter Model	Two-Parameter Model
Mean	100	100
Median	103	102
Minimum	40	54
Maximum	141	146
Percentile Value	**NMI and 95% CI**	**NMI and 95% CI**
20^{th}	85 ± 30	86 ± 30
40^{th}	98 ± 18	96 ± 16
60^{th}	107 ± 13	106 ± 10
80^{th}	117 ± 15	117 ± 9

Clearly, the NMI scale values computed by both one and two parameter models are comparable. However, the latter model produces a marginally narrower confidence interval, particularly in the zone of moderate morbidity (i.e. 60th percentile and above). Tables 4.14 and 4.15 contain the morbidity items in an ascending sequence by their scale values (the difficulty parameter) for the final '50-Item NMI', obtained from both one and two parameter models, respectively. The sequence of items estimated by the one-parameter model seemed more clinically agreeable to the subject-matter experts. For example, the Item 4 (*Apgar score of <7 at 5 minute*) is obviously less morbid than the Item 7 (*Apgar score <7 at 10 minute*), and their scale values by the one-parameter model were 124 and 140, respectively (Table 4-14). Whereas, the scale values of both Item 4 and Item 7 items were estimated at 130, by the two-parameter model. Similarly, the scale values of the Item 49 (Intraventricular Hemorrhage-grade 1 & 2) and the Item 50

(Intraventricular Hemorrhage-grade 3 & 4) estimated by the one-parameter model were 147 and 175, respectively; whereas, using the two-parameter model, their scale values were not considered sufficiently apart at 142 and 145, respectively. Further, the 411 morbidity estimates obtained using the one vs. two-parameter model were compared. The subject-matter experts considered the morbidity estimates obtained using the one-parameter model as better reflections of the impressions of morbidity in the study population. It is possible that increasing the sample size to approximately 15-20 subjects for each item in the tool, i.e. 750-1000 newborns for item analysis using the two parameter model will result in a more stable, reliable and clinically relevant sequence of items in the NMI.

Table 4-14.

Morbidity Items in Ascending Order by Scale Values
One Parameter Model
N = 50 Items

BIN	Morbidity Item	Rescaled Value
30	Poor sucking within 24 hr.	83
19	Apnea detected (by apnea monitor)	96
44	Serum bilirubin >170 micromol/L	97
31	Poor sucking between 24 hr. - 7 days	98
45	Serum bilirubin >250 micromol/L OR Need for phototherapy	103
34	Assisted ventilation within 24 hr.	104
20	Apnea and need for oxygen therapy	113
35	Assisted ventilation beyond first 24 hr.	115
32	Poor sucking beyond day 7	118
9	Heart rate/minute (160-200) (2 or more consecutive readings)	118
4	Apgar score of <7 at 5 minute	123
2	Resuscitation at birth (intubation)	124
16	Hypotonia persisting (or identified) between 1-120 hr.	126
36	Mechanical ventilation within 24 hr.	127
14	Respiratory rate/minute <30 or >60 in 3-24 hr. (2 or more consecutive readings)	128
37	Mechanical ventilation between 24 hr. - 7 days	131
12	Low systolic BP - (2 or more consecutive readings)	132
	(<30 mm Hg for 28-32 weeks) (<40 mm Hg for 32-42 weeks)	
33	Persistent vomiting	132
27	Level of consciuosness - Drowsy/Lethargic	133
24	Tremors/Single seizure	133
21	Apnea and need for resuscitation therapy	135
13	Respiratory rate/minute >100 between 3-24 hr. (2 or more consecutive readings)	136
39	Low urine flow (<2 ml/kg/hour)	136
1	Cord blood pH <7.10	137
7	Apgar score <7 at 10 minute	140
15	Altered colour - Dusky/Central cyanosis	140
22	Thrombocytopenia with or without bleeding disorder	142
	Bleeding - Lungs/Skin/GIT	

Table 4-14.

Morbidity Items in Ascending Order by Scale Values
One Parameter Model
N = 50 Items

BIN	Morbidity Item	Rescaled Value
43	Plasma glucose (<2.2 mmol/L)	146
49	Intraventricular hemorrhage - grade 1 and 2	147
18	Flaccidity present between 1-120 hr.	148
3	Meconium below cords	149
25	More than single seizure	149
17	Hypotonia beyond 120 hr.	150
11	Heart rate/minute <100 within 24 hrs. of birth (2 or more consecutive readings)	151
38	Mechanical ventilation beyond day 7	154
5	Apgar score <4 at 5 minute	155
47	Bacterial culture positive - Blood	155
29	Cardio-pulmonary resuscitation any time before discharge	155
23	Blood transfusion due to Item 22	163
40	Fracture Long Bone/Clavicle/Skull	169
46	Serum bilirubin >340 mmol/L OR Exchange transfusion	171
42	Subdural/Intracerebral bleeding	171
8	Apgar score <4 at 10 minute	172
41	Birth trauma - Nerve injury (Facial/Peripheral)	175
10	Heart rate/minute >200 within 24 hr. (2 or more consecutive readings)	175
26	Seizures and >2 drugs used for treatment	175
50	Intraventricular hemorrhage - grade 3 and 4	175
28	Level of consciousness - Stupor/Obtundation/Coma	179
48	Bacterial culture positive - CSF	179
6	Apgar score <1 at 5 minute	184

Table 4-15.

Morbidity Items in Ascending Order by Scale Values
Two Parameter Model
N = 50 Items

BIN	Morbidity Item	Rescaled Value
30	Poor sucking within 24 hr.	87
44	Serum bilirubin (>170 micromol/L)	94
19	Apnea detected (by apnea monitor)	96
31	Poor sucking between 24 hr. - 7 days	99
45	Serum bilirubin >250 micromol/L OR Need for phototherapy	104
20	Apnea and need for oxygen therapy	111
34	Assisted ventilation within 24 hr.	112
32	Poor sucking beyond day 7	115
35	Assisted ventilation beyond 24 hr.	116
9	Heart rate/minute (160-200) - (2 or more consecutive readings)	121
2	Resuscitation at birth (Intubation)	122
36	Mechanical ventilation within 24 hr.	122
16	Hypotonia persisting (or identified) between 1-120 hr.	122
37	Mechanical ventilation between 24 hr. - 7 days	124
39	Low urine flow (<2 ml/kg/hour)	128
27	Level of consciuosness - Drowsy/Lethargic	129
12	Low systolic BP - (2 or more consecutive readings)	129
	(<30 mm Hg for 28-32 weeks) (<40 mm Hg for 32-42 weeks)	
21	Apnea and need for resuscitation therapy	129
7	Apgar score <7 at 5 minute	130
4	Apgar score of <7 at 10 minute	130
24	Tremors/Single seizure	131
14	Respiratory rate/minute <30 or >60 in 3-24 hr. - (2 or more consecutive readings)	133
25	More than single seizure	134
33	Persistent vomiting	135
29	Cardio-pulmonary resuscitation any time before discharge	136
17	Hypotonia beyond 120 hr. of birth	136
22	Thrombocytopenia with or without bleeding disorder	139
	Bleeding - Lungs/Skin/GIT	

Table 4-15.

Morbidity Items in Ascending Order by Scale Values
Two Parameter Model
N = 50 Items

BIN	Morbidity Item	Rescaled Value
18	Flaccidity present between 1-120 hr.	139
15	Altered colour - Dusky/Central cyanosis	139
38	Mechanical ventilation beyond day 7	142
49	Intraventricular hemorrhage - grade 1 and 2	142
13	Respiratory rate/minute >100 between 3-24 hr. - (2 or more consecutive readings)	144
26	Seizures and >2 drugs used for treatment	145
50	Intraventricular hemorrhage - grade 3 and 4	145
5	Apgar score <4 at 5 minute	146
8	Apgar score <4 at 10 minute	146
1	Cord blood pH <7.10	147
23	Blood transfusion due to Item 22	148
28	Level of consciousness - Stupor/Obtundation/Coma	148
47	Bacterial culture positive - Blood	149
6	Apgar score <1 at 5 minute	153
48	Bacterial culture positive - CSF	166
43	Plasma glucose <2.2 mmol/L	166
11	Heart rate/minute <100 within 24 hr. - (2 or more consecutive readings)	167
40	Fracture Long Bone/Clavicle/Skull	170
42	Subdural/Intracerebral bleeding	172
10	Heart rate/minute >200 within 24 hr. - (2 or more consecutive readings)	179
3	Meconium below cords	186
41	Birth trauma - Nerve injury (Facial/Peripheral)	201
46	Serum bilirubin >340 mmol/L OR Exchange transfusion	211

4.2.2.4 Evaluation of the Degree of Fit of Items on the Scale

The following tests were employed to determine the goodness-of-fit of items on the scale:

(a) Whether the test data satisfied the assumptions of the test model

i) Dimensionality testing

This is discussed in section 4.2.1. There was no violation of assumptions due to this criterion.

ii) Plot of content-based vs. test-based item parameter estimates

The content-based morbidity estimates were obtained following analysis of the information provided by the subject-matter experts during the 'working sessions'. The experts provided weight to each item in the provisional morbidity item-list. The test-based morbidity estimates were obtained subsequent to final item analysis and parameter estimation. Plots of content-based versus test-based morbidity estimates from both one and two parameter models were drawn and are shown in Figures 4-6 and 4-7, respectively.

A correlation coefficient (r) of 0.88 was obtained between the content-based and test-based estimates. A 'Lowess' regression line at 50% fit was drawn between the two estimates of morbidity. The 50% fit attempts to draw a regression line fitting through 50% of the points on the plot. The regression was linear in the moderate and severe morbidity zone. There was no violation in assumptions due to this criterion.

Figure 4-6.

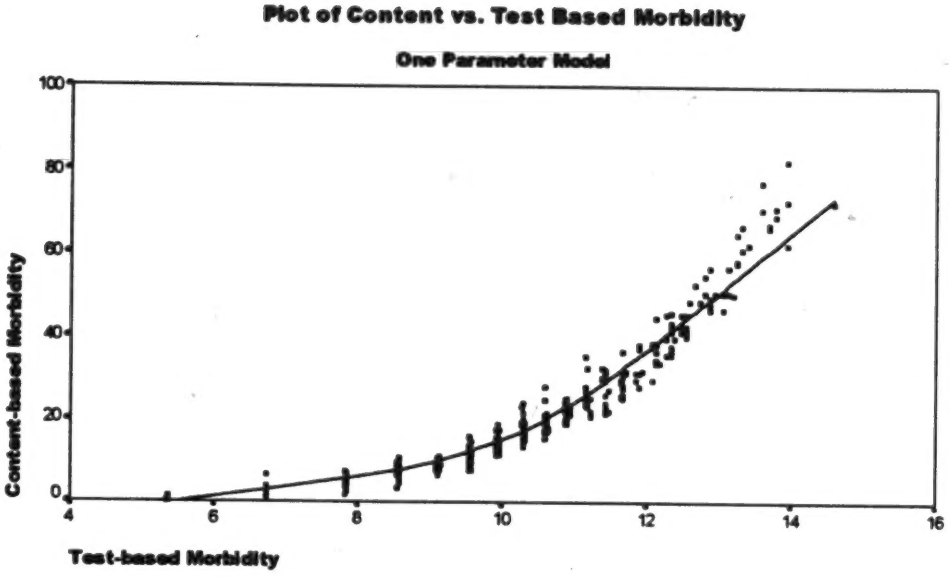
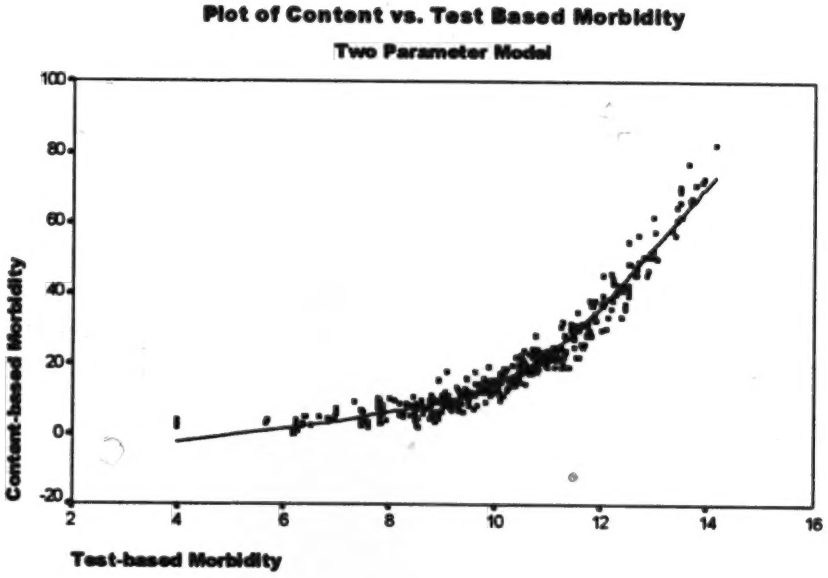


Figure 4-7.



(b) Whether the expected advantages derived from the use of the IRT model were obtained:

The two expected advantages were invariant estimates of *item* and *ability:*

(i) Invariance of the Item Parameter Estimates

For testing this assumption, the study population was subdivided into three groups based on the gestational age, 28 to <32 weeks, 32 to <36 weeks, and 36 to <42 weeks. Of the 411 newborns, 20.8% were between 28 to <32 weeks, 30.5% were between 32 to <36 weeks, and 48.7% were between 36 to <42 weeks gestation. Separate analyses for estimating the item-difficulty parameter were performed for each group. Table 4-16 contains the results of the estimation of the item-difficulty parameter for the three groups obtained using the one-parameter model of IRT.

Table 4-16.

Difficulty Parameter and Gestational Age
Number of Items = 39

	28 to <32 weeks gestation			32 to <36 weeks gestation			36 to <42 weeks gestation		
Serial No.	BIN	Difficulty	Chi-Sq	BIN	Difficulty	Chi-Sq	BIN	Difficulty	Chi-Sq
1	1	3.0	0	1	2.6	0	1	1.5	1.6
2	2	1.1	2.5	2	1.8	3.1	2	1.3	5
3	3	1.5	2.6	4	1.6	4.3	4	1.0	5.7
4	4	4.3	0	5	3.8	0	5	2.9	1.5
5	5	2.7	0	6	2.1	1.1	7	2.4	5.2
6	7	0.3	5.9	7	1.1	4.5	9	1.3	7.3
7	9	3.6	0	8	3.8	0	11	2.6	4.4
8	10	1.4	2.1	9	1.5	3.4	12	2.4	3
9	11	1.8	0.3	10	2.3	1.6	13	2.2	12.9
10	12	1.1	3.1	11	1.8	0.7	14	1.8	4.6
11	13	2.5	0	12	2.6	1	15	2.2	2.7
12	14	1.4	0.4	13	1.8	5.6	16	1.3	5.6
13	15	3.0	0	14	3.0	0	17	2.9	0.9
14	16	2.6	0	15	3.2	0	18	2.8	0.4
15	17	-2.2	0.5	16	-1.0	1.9	19	0.6	2.3
16	18	-0.6	3.6	17	0.5	5.6	20	1.6	2.4
17	19	1.4	4	18	2.5	0.8	21	2.2	3.1
18	20	1.7	2.2	19	3.0	0	22	2.7	2.9
19	22	2.8	0	20	1.9	2.3	24	1.7	6.9
20	23	3.3	0	21	3.3	0	25	2.6	5.1
21	24	4.3	0	22	4.9	0	26	4.4	0
22	25	1.7	2.9	23	2.1	3.2	27	1.9	3.5
23	27	4.3	0	24	3.6	0	29	2.9	0.9
24	28	-2.9	0	25	-1.9	3.2	30	-0.3	4.2
25	29	-2.6	0	26	-0.7	7.3	31	0.9	19.7
26	30	-1.1	5.5	27	0.8	2.9	32	2.6	1.6
27	31	0.9	6.8	28	1.7	8.9	33	2.9	1.1
28	32	-0.3	47.3	29	0.4	36	34	0.3	38.8
29	33	-0.2	4.4	30	0.6	7.9	35	1.6	12.7
30	34	1.1	2.5	31	1.5	0.9	36	1.7	3.9

Table 4-16.

Difficulty Parameter and Gestational Age
Number of Items = 39

	28 to <32 weeks gestation			32 to <36 weeks gestatio			36 to <42 weeks gestation		
Serial No.	BIN	Difficulty	Chi-Sq	BIN	Difficulty	Chi-Sq	BIN	Difficulty	Chi-Sq
31	35	1.4	2.2	32	1.8	5.6	37	2.1	4.9
32	36	2.5	0	33	3.6	0	38	3.6	0
33	37	1.4	2.8	34	2.2	2.7	39	2.5	2.5
34	38	4.3	0	37	4.2	0	42	4.4	0
35	39	2.6	0	38	3.0	0	43	2.6	1.8
36	40	-2.3	0.5	39	-0.8	12.3	44	0.7	30.8
37	41	-1.5	0.7	40	-0.5	14	45	1.1	28
38	42	2.3	0	42	3.3	0	46	4.0	0
39	44	1.7	0.2	44	3.6	0	47	4.4	0

Serial No.= Serial number for items in this table; BIN=Binary Item Number (Refer page 45; Chapter 3); Difficulty=Difficulty Parameter; Chi-Sq.= Chi-Square parameter.

Of the 50 items in the final proposed index of morbidity, only 39 items were available for this analysis since there were no endorsements (scoring) on 11 items in any one of the three categories by gestational age. Table 4-16 contains the estimates using 39 items in each category. Of the 39 items, only 31 items were common to all the three groups. On analysis, the sequence and the scale values of the items on the difficulty scale were quite similar, particularly in the first two groups, i.e. 28 to <32 weeks and 32 to <36 weeks. Table 4-17 contains the results of analysis of correlation (r) of the item-difficulty parameter between the groups 28 to <32 weeks and 32 to <36 weeks; 32 to <36 weeks and 36 to <42 weeks; and 28 to <32 weeks and 36 to <42 weeks.

Table 4-17.

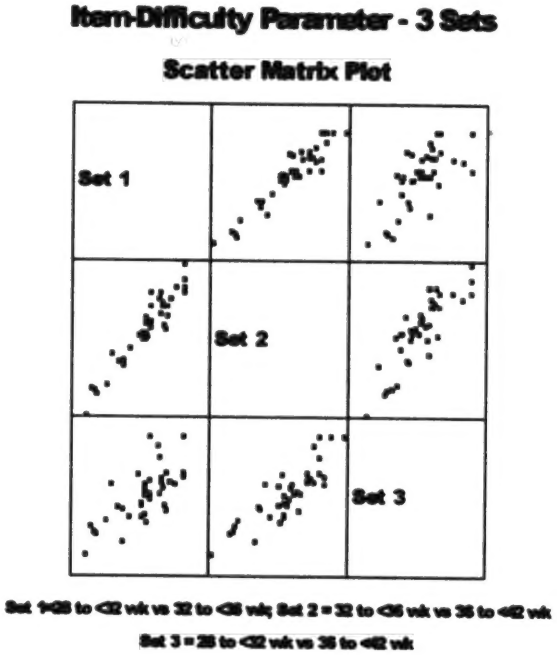
Pearson's Correlation Coefficient: Sets by Gestational Age
Number of Items = 39

	Correlation Coefficient	p-value
(28-<32 and 32-<36) weeks	0.95	.000
(32-<36 and 36-<42) weeks	0.86	.000
(28-<32 and 36-<42) weeks	0.72	.000

Scatter matrix plots of the item-difficulty parameter were drawn between the three groups and are shown in Figure 4-8. The plots revealed that the regression between the groups 28 to <32 weeks versus 32 to <36 weeks was linear, whereas there was a relatively larger degree of scatter in the regression plot of 28 to <32 weeks versus 36 to <42 weeks. This confirmed that the item-difficulty parameters in the group 28 to <32 weeks were similar to those parameters in the group 32 to <36 weeks. However, there was a disparity of this

statistic between the group 36 to <42 weeks versus the other two groups. This disparity among the groups may be due to one or more of the following reasons: differential item functioning due to gestational age, unequal number of subjects in each group, small sample size particularly in the groups 28 to <32 weeks and 32 to <36 weeks, and reduction in the number of items for testing, from 50 to 39.

Figure 4-8.

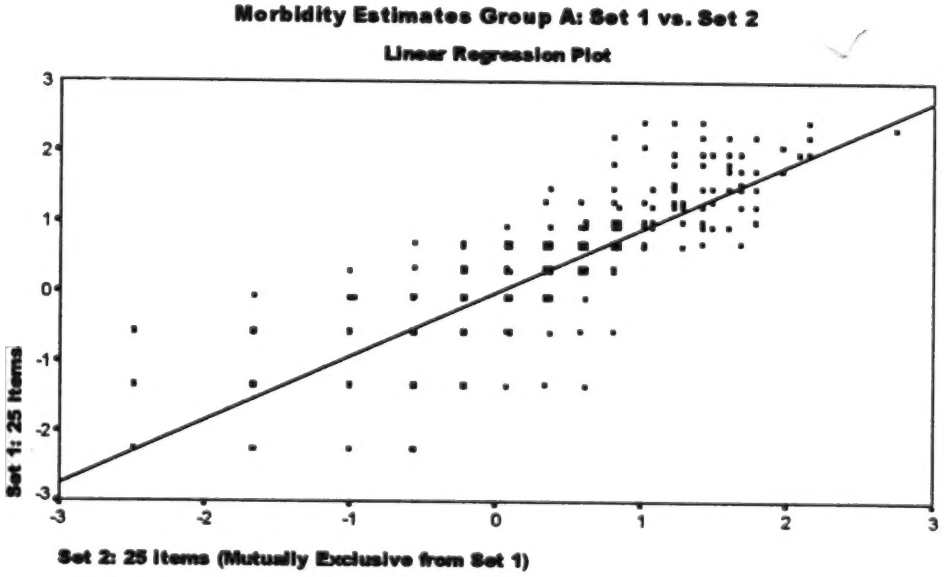


The results provide evidence that there was no serious violation in the goodness-of-fit criterion due to this assumption. However, in the researcher's opinion, the goodness-of-fit criteria of invariance of the difficulty parameter due to gestational age should be tested in future studies including larger sample sizes.

(ii) Invariance of the Latent-Trait i.e. Morbidity Parameter Estimates

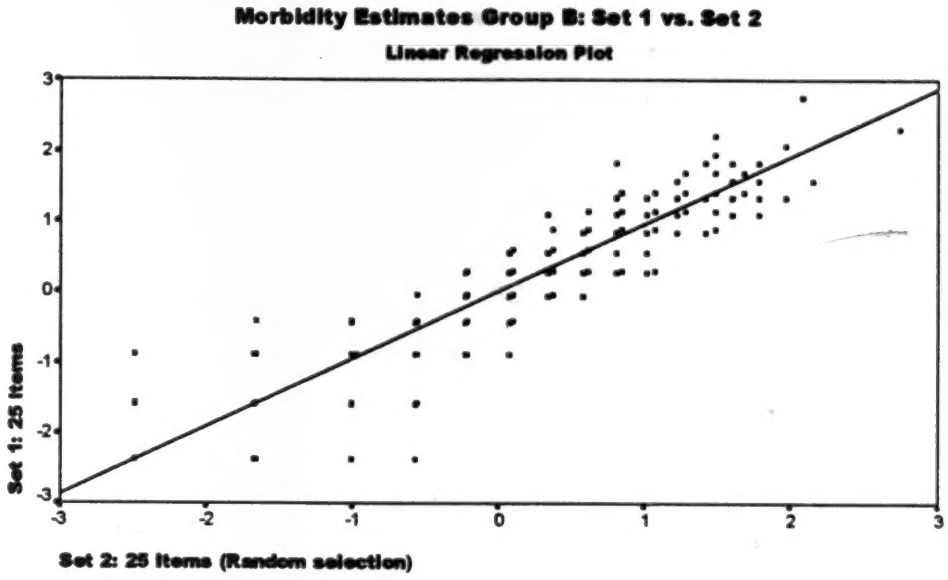
Two groups with two sets of items in each group were obtained from the total item set of 50 items. Group A contained two sets of mutually exclusive 25 items each. Group B contained two sets of randomly selected 25 items each. Morbidity estimates were obtained and compared among the two sets (of 25 each) of the total items in each group A and B. A correlation coefficient of 0.89 and 0.94 was computed between the two sets for group A and B, respectively. Figures 4-9 and 4-10 depict linear regression plots of morbidity estimates obtained between set 1 and set 2 of the each group, A and B.

Figure 4-9.



Figures 4-9 and 4-10 provide evidence that the relationship between set 1 and set 2 in both the groups was fairly linear. However, there was an expected degree of scatter in the regression plot in the mild morbidity zone. There was no violation of assumption of the goodness-of-fit criteria due to this criterion.

Figure 4-10.



(c) Whether there was a closeness of fit between predictions and observable outcomes

This implies closeness between the estimated test scores and the raw test data. Statistical tests of the goodness-of-fit of various IRT models have been given by many authors [88-90]. We applied the following formal tests of the goodness-of-fit model:

(i) Chi-Square

A comparison of the observed and predicted test score distributions by the chi-square statistic was performed. A significantly large value of the statistic indicated a failure of fit (p-value >0.05) of one or more of the response models for the n items.

(ii) Standardized Posterior Residuals

Investigations of Residuals and Standardized Posterior Residuals of model-test data fits at the item level were carried out. Standardized Posterior Residuals are the differences between the posterior probability of correct response at selected values of θ and the probabilities at those points computed from the corresponding fitted response model. Each of the seven items that were deleted from the 58-item index had high standardized residuals of the order of 3.5 and above.

(iii) Plots of Test Scores and Ability Estimates

While plotting, a likelihood-ratio chi-square statistic was used to compare the grouped frequencies of endorsed and non-endorsed responses in the intervals with those expected from the fitted model at the interval mean. To diagnose cases of poor fit, inspection of the plot of actual vs. expected frequencies of endorsed and non-endorsed responses was done for the total test as well as for each of the 58 items (refer Appendix 10 and 11).

4.2.2.5 Scoring System

Following the determination of item sequence and scale values on the morbidity continuum, it was recommended to explore a pragmatic and simple scoring system for the items to facilitate field application of the tool. The scores were derived from the scale values of the NMI, according to three different scoring systems:

(a) Endorsed or scored items were summed for each newborn, giving a total number of 'items endorsed' (range 1 to 50).

Two other types of NMI scores by the 'milestone' method were computed for each newborn. The scale values of morbidity items as sequenced by the 'difficulty scale' (Table 4-14 and 4-15) were used for computing these scores as follows:

(b) The first was obtained by aggregating the scale values of the highest three items scored by each newborn.

(c) The second was obtained by aggregating the scale values of the highest five items scored by each newborn.

In the study population, 67.2% newborns scored on five or more items and 83.2% scored on three or more items. For newborns scoring on less than three items, the scale values of the items scored were aggregated for this analysis. The latter scores, (b) and (c), were computed using the scale values from both one and two parameter models. The correlations of the three types of scores, (a), (b), and (c) with the NMI scale values were

computed for all the 411 newborns. Table 4-18 contains the results of this analysis. The correlations of all the scoring systems with the NMI values were uniformly high.

Table 4-18.

Correlation between NMI and Scoring Systems

NMI and Scoring System	One Parameter (r)	Two Parameter (r)
NMI and Scored Item (a)	0.93	0.90
NMI and Three-Item Score (b)	0.91	0.88
NMI and Five-Item Score (c)	0.96	0.91

(r) statistically significant at $p < 0.05$ for all the six values

The values of the Pearson's correlation coefficient (*r*) did not provide any concrete evidence regarding the choice of one model over the other. However, the correlation between the Five-Item Score (c) and the NMI scale values from the one and the two-parameter models at 0.96 and 0.91, respectively, was the highest. Considering that the sequence of items by the one-parameter model was more acceptable to the subject-matter experts, we elected to use the Five-Item Score derived from the one-parameter model for subsequent analyses of external validation. It is recommended that definite judgments regarding the choice of one model over the other should be based on the results of field-testing of the item scale values derived from both models.

4.2.3 External Validation

External validation involved determining the correlation between the 'Five-Item Score' obtained from the 'one parameter model' and external criterion measures of both cross-sectional and longitudinal events. The collected data on the external criterion measures were either continuous or categorical.

4.2.3.1 External Validation: Continuous Measures

For the continuous measures, Pearson's correlation coefficient, *r* was computed between the Five-Item Score and the test variable on the continuous scale. Table 4-19 contains the results of correlations of the Five-Item Score with birth weight, gestational age, SNAP and NTISS scores, physician's concurrent assessment, length of stay and postpartum nursing utilization.

In accordance with the accepted view that low birthweight and low gestational age predispose newborns to increased morbidity at birth, a statistically significant negative r of the Five-Item Score with the birth weight and the gestational age (-0.38 and -0.33) was obtained. The finding suggested that the lower the birth weight or gestational age, the higher the morbidity score. For concurrent validation, the ratings by the attending physicians on the analogue scale for day 1, 3, 7 and/or discharge were summed together, and the correlation of the summed assessment score on the analogue scale with the Five-Item Score was computed. The *r* with the summed score was 0.69 at $p < 0.05$, and the *r* with the individual ratings on day 1, day 3 and day 7 was 0.70, 0.63 and 0.55,

respectively, at $p < 0.05$. High correlations between the physician's assessments and the Five-Item Score indicate the usefulness of an extensive literature review and an appropriate choice of items by the subject-matter experts during the substantive validation phase. The scaling and subsequent scoring of the morbidity index lends objectivity to subjective assessments made by the subject-matter experts, while assessing the severity of morbidity in newborns.

Table 4-19.

External Validation Events: Continuous Variables

External Validation Events	N	Pearson's Correlation (r)	p-value
Cross-Sectional Events			
Birth Weight	411	-0.38	.000
Gestational Age	411	-0.33	.000
Concurrent Physician's Assessment	165	0.69	.000
SNAP Score	248	0.53	.000
NTISS Score	139	0.59	.000
Longitudinal Events			
Total Length of Stay	411	0.55	.000
LOS (A+B side)	393	0.47	.000
LOS (C side)	299	0.40	.000
Postpartum Utilization			
Time Spent	87	0.30	0.002
Contacts	87	0.23	0.03

Abbreviations: SNAP = Score for Neonatal Acute Physiology; NTISS = Neonatal Therapeutic Intensity Sub Score; LOS = Length of Stay.

The SNAP for a newborn was available at 24 hours, 3 days, and 14 days of age. We summed the three SNAP scores for each newborn to correspond with the NMI scale values obtained from assessments conducted over a 7-day period. The NTISS is a score for therapeutic intensity over the entire length of stay. The correlation of the Five-Item Score with the SNAP and the NTISS was 0.53 and 0.59, respectively, at $p<0.05$.

Among the longitudinal events, there was a statistically significant r of 0.55 between the Five-Item Score and the total Length of Stay (LOS) in the hospital. Newborns with low birth weight and *no other significant morbidity* are admitted to the step-down nursery (C side) for extended periods, until such time as they start gaining weight and eventually acquire a weight of 2000g. Accordingly, an r of 0.40 at $p<0.05$ was obtained between the Five-Item Score and the LOS at the C side NICU, whereas an r of 0.55 and 0.47 was obtained with the total LOS and the LOS at the A+B side, respectively. For correlation with the postpartum utilization, the 'time spent' was computed as the duration of time (in minutes) spent on all the phone calls made to the family for infants' needs. 'Contacts' was computed as the sum of the number of phone calls and the number of visits (one visit was computed as equivalent to two phone calls) made to the family for infants' needs. A correlation of 0.32 at $p<0.05$ between the Five-Item Score and the 'time spent' was obtained. Since contacts with the family by the PHN often served multiple purposes, it was extremely difficult to categorically state the purpose of phone calls or visits from the charted nurse's notes. The purpose of the postpartum program is to provide social, psychological and emotional support in addition to medical support. As such, it is probably not appropriate to use this program as one of the criteria for external validation.

4.2.3.2 Distribution of the Five-Item Score of Morbidity

Table 4-20 contains the results of descriptive statistics of the Five-Item Score for 411 newborns. As mentioned earlier, the Five-Item Score was computed by summing the scale values of the 'difficulty parameter' of the five highest items scored by a newborn on the morbidity index. The Five-Item Score for the newborns recruited from the regular nursery ranged from 0 to approximately 200. The score range for the newborns admitted to the NICU for observation (only), was higher than the regular nursery newborns, ranging from approximately 80 to 300. The average value of the Five-Item Score for the 411 newborns was 497, ranging between 0 to 836. The standard error of estimate and the confidence interval around the score were computed by methods similar to those used for the total NMI, depicted in Table 4-13. The confidence interval for the scores in the middle portion of the Five-Item Score was ± 30.

Table 4-20.

Descriptive Statistics: Five-Item Score

Statistics	Five-Item Score
Mean	497
Median	567
Minimum	0
Maximum	836
Percentile Value	
20^{th}	315
40^{th}	520
60^{th}	598
80^{th}	674

4.2.3.3 External Validation: Categorical Measures

There were four categorical external validation events in our data set including neonatal mortality, fetal distress, small for gestational age (SGA) and maternal disease conditions. We considered the diagnosis of fetal distress only when the observations of prolonged decelerations and loss or decrease in fetal heart variability were made by attending physicians. For the diagnosis of SGA, we confirmed it with the help of gender-specific 'norming' plots between birth weight and gestational age, available at the Perinatal Clinic (Arbuckle et al, 1985). The mean values of the categorical variables on the Five-item score were compared using the independent sample t-test.

Table 4-21.

External Validation Events: Categorical Variables
Independent Sample t-test

External Event	Mean Difference in the Five-Item Score	95% CI for Difference	2-Tail Significance
Expired vs. Alive (n=6 vs. 405)	260	(204, 319)	0.00
Fetal distress vs. no Fetal distress (n=65 vs. 346)	98	(52, 144)	0.00
SGA vs. Non-SGA (n=63 vs. 348)	65	(15, 115)	0.01
Maternal age ≥40 years vs. <40 years (n=10 vs. 401)	11	(-122, 144)	0.87

Table 4-21 contains the results of an independent sample t-test for the comparison of means of the Five-Item Score between the 'expired versus alive', 'SGA versus non-SGA', 'fetal distress versus no fetal distress' and 'maternal age <40 years versus ≥40 years'. The results in the Table 4-21 agree with the accredited relationship between neonatal mortality, fetal distress and diagnosis of SGA, and morbidity at birth. We found no statistically significant relationship between morbidity at birth and maternal age. Literature provides conflicting opinions regarding a definite cut-off level of maternal age with regards to increased adverse outcomes in newborns. However, analyzing maternal age as a continuous variable also did not produce any significant correlation between the maternal age and the Five Item Score.

4.2.3.4 Representative Population and Maternal Events

Approximately 15-20% of the newborns in the general population have some degree of morbidity at birth. We wanted to compute an estimate of the average Five-Item Score for the general population. To accomplish this task, we collected a subset of 50 newborns from the 411 recruited for the study. This subset consisted of relatively healthy and morbid newborns in the ratio of 4:1. Therefore, 40 newborns in this subset were from those admitted to the regular nursery or the NICU for observation only, and 10 newborns were a random sample of the rest of the population admitted to the NICU. We named this subset as 'representative subset', and the average Five-Item Score for newborns in this subset was found to be 300 (± 95% CI of 60).

From our data set, we computed the mean values of the Five-Item Score for the maternal disease states mentioned in Appendix 8. Table 4-22 contains the results of these analyses for the given population of 411 (morbid) newborns. For these maternal variables, the tests of statistical significance were not recommended because of two important considerations that could lead to biased conclusions. First, although we obtained an average population estimate of the Five-Item Score from the 'representative subset', this estimate was not based on a truly generalizable 'norm' group of newborns, and second, in the given population, there were multiple influences other than the said maternal disease conditions

Table 4-22.

Five-Item Score and Maternal Disease Conditions

Average Five Item Score for the Recruited Population, (n=411) = 497 (± 30)
Average Five Item Score for the Representative Population, (n=50) = 300 (±60)

Maternal Conditions	Five-Item Score
Pregestational Diabetes Mellitus, (n=10)	558
Antepartum Hemorrhage, (n=26)	618
HELLP Syndrome, (n=6)	573
Chorioamnionitis, (n=5)	522
Oligohydroamnios, (n=29)	540
Gestational Diabetes Mellitus, (n=15) (controlled with diet)	484
Gestational Diabetes Mellitus, (n=15) (controlled with insulin)	450
Obesity, (n=43) (pre-pregnancy weight >90 kg)	456
Diagnosis of GBS, (n=39)	435

that were contributing to newborn morbidity, thus making a newborn eligible for admission to the NICU. For instance, a newborn of a mother with gestational diabetes mellitus will not be admitted to the NICU unless warranted by some considerations in addition to this maternal condition. Therefore, the comparison between the estimated population average (from the representative subset) and the average due to maternal events in the given population are not truly consequential. However, the Table 4-22 provides an essence of the gradient in score values that could result due to various maternal conditions. The subject-matter experts concluded that the gradient in these scores due to maternal disease conditions was clinically relevant and agreed with the expected neonatal outcomes in these conditions of maternal morbidity.

Based on the results of the substantive, structural and external validation, a newborn morbidity index consisting of 50 morbidity items is proposed. Appendix 12 contains the item-list and the corresponding binary item numbers (column 'Check') for the proposed outcome measure, NMI. Table 4-14 contains the scale values for the 50 binary items of morbidity in the NMI, furnished by the one-parameter model of the IRT. Of the scoring systems, the Five-Item Score is recommended for application in both clinical and epidemiological studies pending further refinements of the NMI following field-testing.

Chapter 5

Summation, Utilization, Strengths, Limitations and Future Implications

5.1 Summation

This research project involved construction and validation of an obstetrical outcome measure, Newborn Morbidity Index for use as a discriminative tool to measure alteration in neonatal outcome between alternate obstetrical therapeutic approaches or maternal conditions. A sample of 411 newborn subjects, having very mild to moderately severe degrees of morbidity at birth and admitted to the nurseries at the Royal Alexandra Hospital in Edmonton was recruited. The research project achieved its aims. A sensitive, reliable, discriminative outcome measure for obstetrical studies - *Newborn Morbidity Index* was developed, and substantively, structurally and externally validated. Recommendations for the scaling and the scoring models were made.

To our knowledge, this is the first detailed project to develop a discriminative obstetrical outcome measure utilizing sound psychometric techniques. Despite acknowledgment in the existing literature of a lack of standardized indicators or outcome measures following obstetrical studies, previously developed tools meet the criteria of substantive validation only, and have not been standardized or further evaluated with respect to their psychometric properties.

The proposed morbidity index is focused on morbidity in the mild to moderate range, the

range relevant to most obstetrical studies that propose to examine conditions other than preterm labor and delivery. Since this was a scaling study, a primary criterion for recruitment was a uniform representation of all grades of morbidity. The origin of morbidity, diagnosis or maternal confounding factors were not taken into consideration. By the same token, the recruited group of newborns did not represent a random sample of the general population. For the purpose of application of the developed morbidity score to population studies, a 'generalizable' norming criterion will need to be established in future.

Investigation of the structure of the NMI using DIMTEST program for dimensionality testing demonstrated that it is a unidimensional index and that the single dimension underlying the item scores is newborn morbidity. The latent trait model, IRT provided a useful framework for determining the indices and sequencing of parameters in the data set. The BILOG program for parameter estimation and scaling provided effective analytic technique for this index. The 'Maximum Likelihood Estimation' technique for analysis furnished stable and robust parameter estimation, and adequate goodness-of-fit criteria for the data set.

Regarding sequencing of items in the index, the sequence suggested by the one-parameter model in contrast to the two-parameter model seemed more relevant with regards to clinical judgment. Though fitting indices with the two-parameter model resulted in greater 'information' and narrower 'confidence bands', particularly in the moderate to severe morbidity zone, a larger sample size (approximately one thousand subjects) will

likely furnish a more stable, robust and clinically sound sequence and scale values of the morbidity items. However, the best judgment towards the adequacy of one model over the other rests on the following two factors: firstly, the opinion of the subject-matter experts, and secondly, the field-testing of the index values from both models in various maternal conditions, settings and situations. In keeping with the results of the analysis for the most suitable scoring model, it is recommended that the Five-Item Score be employed for application of the Newborn Morbidity Index. Situations or studies examining very mild degree of morbidity in newborns may employ the Three-Item Score since the correlations between the Three-Item Score and the Newborn Morbidity Index were quite close to the correlations between the Five-Item Score and the Newborn Morbidity Index ($r = 0.96$ versus 0.91).

An interesting observation made while exploring the structure of newborn morbidity was that there is an area of very mild to mild morbidity, which may be labeled as a 'gray zone' of morbidity. A careful examination of the morbidity items from this zone revealed that they were not internally consistent with the concept of morbidity. They neither furnished 'information' nor 'discrimination' attributes to the scale values. The sole contribution of those items to the scaling criteria was an increase in measurement error or (so-called) "noise" but not the true "signal" or consistency in the measure. Obstetrical studies that incorporate those individual items as end-points or outcome variables might overlook the target, the underlying trait of morbidity. A few examples of such items are: *bag and mask at birth, assisted ventilation or poor sucking in the first 24 hours, peripheral cyanosis, serum bilirubin between 170-250 mmol/L, meconium above cords,*

hypotonia in the first 1 hour, and *hyperalert/hyperirritable state*. In the future studies, it may be worthwhile to remove a few other items containing low information from the proposed index, e.g. *oral feeding difficulty in the first 24 hours, assisted ventilation in the first 24 hours*, without compromising the utility or merit of the NMI.

5.2 Utilization

The following summarizes the properties, salient features, performance characteristics, and further plans for utilization of the Newborn Morbidity Index (NMI).

5.2.1 Properties

(a) Nature and Purpose: It is a discriminative outcome measure for use in the field of Maternal-Child Health to compare therapeutic strategies and maternal disease conditions in clinical trials, clinical comparative and epidemiological studies. For instance, to examine the influence of the following conditions on outcomes in newborns at birth: antepartum events, such as pregnancy induced hypertension, diabetes mellitus, maternal infections, maternal chronic diseases, maternal age, maternal drug addiction, maternal smoking etc.; intrapartum events, such as prolonged labor, operative delivery, breech vaginal delivery, abruptio placentae, fetal distress, meconium aspiration etc.; fetal conditions, such as asphyxia, fetal distress, small for gestational, and post-maturity etc.

(b) Recommended population: It is recommended that the NMI be used in maternal or fetal conditions leading to mild to moderately severe degree of morbidity in newborns at $\geq$28 weeks gestation.

(c) Recommended model for scaling and scoring: One-parameter model of the IRT is recommended for item analysis and item calibration of the NMI consisting of 50 binary morbidity items. The recommended model for scoring is the milestone model aggregating the highest five items scored on the NMI, the Five-Item Scoring System.

5.2.2 Salient Features

(a) Number of items: The total number of binary items of morbidity in the index is 50.

(b) Data collection form: Data collection form is provided in Appendix 12. Data collectors will be required to check mark the items scored by a newborn in the column 'Check'. The numbers in this column correspond to the 'binary item number' used for item analysis.

(c) Scale values: Table 4-14 contains the recommended scale values for computing morbidity index.

(d) Data extraction: Documented information on the morbidity items may be collected either prospectively or retrospectively from the newborn's chart.

(e) Data collection: For prospective data collection, the data may be charted at the end of day 1, 3 and 7 (or discharge, whichever is earlier).

(f) Recommended hospital setting: Provision of a secondary level care (or above) nursery offering standard newborn examination by pediatrician(s) with a facility for routine laboratory tests including cord blood pH, plasma glucose, serum bilirubin, platelet count and bacterial culture is required.

5.2.3 Performance Characteristics

(a) Time required for data collection: Depending on the complexity of morbidity, 10 to 20 minutes per newborn chart are required for data collection.

(b) Reliability coefficient: A reliability coefficient of 0.86 was computed for the NMI.

(c) Estimated 95% confidence band of measurement error for the Five-Item Score:

Range of Five-Item Score	95% Confidence Band (± 1.96 X SEE)	Range of Morbidity
< 300	Mean ± 25% (of the mean)	Very Mild
300-499	Mean ± 20% (of the mean)	Mild
500-749	Mean ± 10% (of the mean)	Mild-Moderate
>750	Mean ± 20% (of the mean)	Severe

5.2.4 Future Plan for Utilization

The plan is to (a) prepare an electronic form for marking items on the NMI and the Five-Item Score, (b) apply the NMI scale values in obstetrical clinical studies and (c) 'norm' the NMI in the general population for use in population-based epidemiological studies

5.3 Strengths and Limitations

5.3.1 Strengths

As recommended, the NMI is easily administered, cost-efficient, sensitive, valid, reliable, non-laboratory intensive, and generalizable to various maternal conditions, settings, and situations. There are no available standardized obstetrical outcome measures designed to evaluate the mild to moderate range of morbidity. This range is particularly interesting to most obstetrical studies that intend to examine conditions other than preterm labor and delivery.

5.3.2 Limitations

The NMI as developed has been validated only for the mild to moderate degree of morbidity in newborns at $\geq$ 28 weeks gestation. The test information function curve (Figure 4-2) indicates that it will serve to discriminate better in the mild to moderate range of morbidity than in the very mild or the severe range of morbidity. Clearly, it is not recommended for the studies of maternal conditions that (possibly) result in preterm delivery.

5.4 Future Implications

5.4.1 Pertaining to Design

A larger sample size coupled with closer collaboration with the pediatricians will further confirm the robustness of the findings. Increasing the sample size to 1000 subjects to explore the possibility of using the parameter estimates from the two-parameter model of the IRT.

5.4.2 Pertaining to Items

(a) The index may be modified in future studies to include additional items that are either relevant to the severe degree of morbidity or the very mild degree of morbidity, or both. The feasibility of including items of neurobiological and neurobehavioral states in the domain of newborn morbidity may be explored.

(b) To make the index more responsive, the attributes of 'oral feeding difficulties' (attribute Q, Appendix 7) and 'respiratory status' (attribute R, Appendix 7) may each contain additional categories, i.e. morbidity lasting from 24 hrs.-3 days, and from 4-7 days, in place of morbidity lasting from 24 hrs.-7 days.

(c) Events such as 'surgery at birth', 'congenital anomalies' or 'metabolic abnormalities' should to be categorically defined and have their role examined by a committee of experts.

5.4.3 Pertaining to External Validation

Further validation of the index with the external criteria of 'readmission rate', 'number of visits to the pediatrician's office' and 'developmental assessment' of the infant at 12 and 18 months is proposed.

5.4.4 Pertaining to Psychometrics and Applied Measurement

It may be possible to investigate the structure of the underlying trait of morbidity using other models for item-analysis with IRT and dimensionality testing, and to compare the results of analyses. It will be worthwhile to create an item-bank of morbidity items for further expansion or modification of the NMI. The aspect of 'differential item functioning' for newborns with low gestational age or birth weight needs exploring.

Artificial neural networks employed for making clinical decisions using items of laboratory testing and predictors of disease conditions are based on similar principles as the IRT. It may be beneficial to explore relationships between the two models for application in the field of development of outcome measures.

References

1. Statistics Canada: Health Reports: Births 1990; 1992: Supplement No. 14: 4:1.

2. The Health of Canada's Children: A Canadian Institute of Child Health Profile, 2nd Edition 1994.

3. Amini SB, Dierker LJ, Catalano PM, Ashmead CG, Mann LI. Trends in an obstetric patient population: An eighteen year study. Am J Obstet Gynecol 1994;171:1014-21.

4. Hansen JP. Older maternal age and pregnancy outcome: A review of the literature. Obstet Gynecol Surv 1986; 41:726.

5. Ventura SJ. Trends in birth to older mothers, 1970-79. National Centre for Health Statistics Report 31 (Suppl) 1982.

6. Davidson EC Jr, Fukushima T. The age extremes for reproduction: Current implications for policy change. Am J Obstet Gynecol 1985;152:467.

7. Gabbe SG. A story of two miracles: The impact of the discovery of insulin on pregnancy in women with diabetes mellitus. Obstet Gynecol 1992; 79: 295-299.

8. Apgar VA. A proposal for a new method of evaluation of the newborn infant. Anesth Analg 1953; 32: 260-267.

9. Schmidt B, Kirpalani H, Rosenbaum P, Cadman D. Strengths and limitations of the Apgar score: A critical appraisal. J Clin Epidemiol 1988; 42: 843-850.

10. Jepson HA, Talashek ML and Tichy AM. The Apgar Score: Evolution, limitations, and scoring guidelines. Birth 1991;18: 2: 83-92.

11. Marrin M and Paes BA. Review: Birth Asphyxia: Does the Apgar score have a diagnostic value? Obstet Gynecol 1988; 72:1:120-123.

12. Towell ME. The influence of labor on the fetus and the newborn. Pediatr Clin North

Am 1963; 10: 575-595.

13. Desmond MM, Rudolph AJ, and Phitaksphraiwan P. The transitional care nursery: A mechanism for preventive care in the newborn. Pediatr Clin North Am 1966; 13: 651-668.

14. Brisson-Carroll G, Fraser W, Breart G, Krauss I, Thornton J. The effect of routine early amniotomy on spontaneous labour: a meta-analysis. Obstet Gynecol 1996; 87: 891-896.

15. Yuksel B, Greenough A, Dobson P, Nicolaidas KH. Advanced maternal age and smoking: risk factors for admission to a neonatal intensive care unit. J Perinatal Med; 1996; 24 (4): 397-403.

16. Fiser DH. Outcome evaluations as measures of quality in pediatric intensive care. Pediatr Clin North Am 1994; 4(6): 1423-1438.

17. Van der Ven, AHGS. Evaluation. In: Introduction to Scaling. New York: Wiley & Sons 1980; 263-75.

18. Davies PM, Coxon APM. Key texts in multidimensional scaling. London, England: Heinemann Educational Books;1982.

19. Desmond MM, Franklin R, Valbona C et al. The clinical behavior of the newly born: The term baby. J of Pediatr 1963; 62: 307-325.

20. Drage JS, Kennedy C, Berendes H et al. The Apgar Score as an index of infant morbidity: A report from the Collaborative Study of Cerebral Palsy. Dev Med Child Neurol 1966; 8: 141-148.

21. Drage JS, Kennedy C and Schwarz BK. The Apgar Score as an indicator of neonatal mortality. Obstet Gynecol 1964; 24: 222.

22. Apgar V, James LS. Further observations on the Newborn Scoring System. Am J Dis Child 1962; 104: 419.

23. Towell ME. The infuence of labor on the fetus and the newborn. Pediatr Clin North Am 1963; 10: 575-595.

24. Desmond MM, Franklin R, Blattner R and Hill RM. The relation of maternal disease to fetal and neonatal morbidity and mortality. Pediatr Clin North Am 1961; 8: 421-440.

25. Knaus WA, Zimmerman JE, Wagner DP, et el. APACHE- acute physiologically based classification system. Crit Care Med 1981; 9: 591-597.

26. Neonatal risk factors. In: Neonatal-Perinatal Medicine: Diseases of the fetus and infant. Avroy A. Fanaroff, Richardson J. Martin. Mosby Year Book 1992; (1): 3-4.

27. Kurczymanski TW. Congenital malformations. In: Neonatal-Perinatal Medicine: Diseases of the Fetus and Infant. Avroy A. Fanaroff, Richardson J. Martin. Mosby Year Book 1992; (1): 372.

28. Zinn AB. Inborn errors of metabolism. In: Neonatal-Perinatal Medicine: Diseases of the Fetus and Infant. Avroy A. Fanaroff, Richardson J. Martin. Mosby Year Book 1992; (1): 1118-1150.

29. Scheiner AP, Sexton ME. Prediction of developmental outcome using a perinatal risk inventory. Pediatrics 1991; 88: 1135-1143.

30. Minde K, Whitelaw A, Brown J. Effect of neonatal complications in premature infants on early parent-infant interactions. Dev Med Child Neurol 1983; 25: 763-777.

31. Littman B, Parmelee AH. Medical correlates of infant development. Pediatrics 1978; 61: 470-474.

32. Salmey A, Davis S, Eldredge L, et al. Neonatal status: an objective scoring method

for identifying infants at risk for poor outcome. Early Hum Dev 1988; 17: 233-243.

33. Hobel CJ, Hyvarinen MA, Okada DM, Oh W. Prenatal and intrapartum high risk screening. Prediction of the high risk neonate. Am J Obstet Gynecol 1973; 117: 1-9.

34. Nesbitt REL, Aubry RH. High risk obstetrics II; the value of a semi-objective grading system in identifying the vulnerable group. Am J Obstet Gynecol 1969; 103: 972-976.

35. Aubrey RH, Pennington JC. Identification and evaluation of high-risk pregnancy: the perinatal concept. Clinic Obstet Gynecol 1973; 16: 3-12.

36. Sarnat HB, Sarnat MS. Neonatal encephalopathy following fetal distress. A clinical and electroencephalographic study. Arch Neurol 1976; 33: 696-705.

37. Bayley N. Bayley scales of infant development: Manual 2nd Edition. New York: Psychological Corporation; Harcourt Brace, San Antonio, Tex, 1993.

38. Brazy JE, Eckerman CO, Oehler JM. Nursery Neurobiological Risk Score: Important factor in predicting outcome in very low birth weight infants. J Pediatr 1991; 118: 783-792.

39. Cullen DJ, Civetta JM, Briggs BA, et al. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Crit Care Med 1974; 2: 57-60.

40. Keene AR, Civetta JM, Briggs BA. Therapeutic intervention scoring system: update. Crit Care Med 1983; 11: 1-3.

41. Richardson DK, Gray JE, McCormick MC, Workman K, Goldman DA. Score for neonatal acute physiology: A physiology severity index for neonatal intensive care. Pediatrics 1993; 91: 617-623.

42. Richardson DK, Phibbs CS, Gray JE, McCormik MC, Workman-Daniels K, Goldman DA. Birth weight and illness severity: Independent predictors of Neonatal Mortality.

Pediatrics 1993; 91: 969-975.

43. de-Courcey, Wheeler RH, Wolfe CD, Fitzgerald A, Spencer M, Goodman JD, Gamsu HR. Use of the CRIB (Clinical Risk Index for Babies) Score in prediction of neonatal mortality and morbidity. Arch Dis Child Fetal Neonatal Ed. 1995; 73(1): F32-36.

44. Yeh TS, Pollack MM, Ruttiman UE, et al. Validation of a physiological stability index for use in critically ill infants and children. Peditr Res 1984; 18: 445-451.

45. Phibbs CS, Richardson DK, Gray JE, McCormick MC, Workman, Goldmann D. The relationship between the score for neonatal acute physiology and the charges and length of stay in the neonatal intensive care unit. Clin Res 1992; 40: 98A.

46. Hannah ME, Hannah WJ, Hellmann J, Hewson S, Milner R, Willan A, and the Canadian Multicentre Postterm Pregnancy Trial Group. Induction of labor as compared with serial antenatal monitoring in post-term pregnancy: A randomized controlled trial. N Eng J Med 1992; 326: 1587-92.

47. Fraser W, Marcoux S. Comparing a policy of early vs. late pushing in women with epidurals. 1995-1996; Unpublished sources.

48. Gross PA, Stein MR, Van Antwerpen C, et al. Comparison of severity of illness indicators in an intensive care unit. Arch Intern Med 1991; 151: 2201-2205.

49. Gonnella JS, Hornbrook MC, Louis DZ. Staging of disease: a case-mix measurement. JAMA 1984; 251: 637-644.

50. Medicare Program. Prospective payments for Medicare inpatient hospital services. DHHS, Health Care Financing Administration 42 CFR; Parts 405, 409, 489. Federal Register 1983;48:39752-39890.

51. Baker SP, O'Neill B, Hadon W, et al. The injury-severity score: a method for

describing patients with multiple injuries and evaluating emergency care. J Trauma 1974; 14: 189-196.

52. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974; 2: 81-84.

53. Wheeler JRC, Van Harrison R, Wolfe RA, Payne BC. The effects of burns severity and institutional differences on the cost of care. Med Care1983; 21: 1192-1203.

54. Goldman L, Weinberg M, Weisberg M, et al. A computer driven protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Eng J Med 1982; 307: 588-596.

55. Teres D, Lemeshow S, Avrunin JS, Pastides H. Validation of the mortality prediction model for ICU patients. Crit Care Med 1987;15:208-213.

56. Cullen DJ, Civetta JM, Briggs BA, et al. Therapeutic intervention scoring system: a method for quantitative comparison of patient care. Crit Care Med 1974; 2: 57-60.

57. Keene AR, Civetta JM, Briggs BA et al. Therapeutic intervention scoring system: update 1983; Crit Care Med 1983; 11: 1-3.

58. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of intensive care in major medical centers. Ann Intern Med 1986; 104: 410-418.

59. Wagner DP, Knaus WA, Draper EA, et al. Statistical validation of a severity of illness measure. Am J Public Health 1983; 73: 878-884.

60. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991; 100: 1619-1636.

61. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score.

Crit Care Med 1988; 16: 1110-1116.

62. Piper MC, Pinnell L, Darrah J, Maguire TO & Byrne PJ. Construction and validation of the Alberta Infant Motor Scale (AIMS). Can J of Public Health 1992; 83 (supplement 2): 46-50.

63. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis; 1985; 38: 1: 27-36.

64. Loevinger J. Objective tests as instruments of psychological theory. In: Jackson DN and Messick (Eds.), Problems in Human Assessment. New York: McGraw Hill, 1967: 78-120.

65. Hambleton RK, Cook L. Latent trait models and their use in the analysis of education test data. J of Educational Measurement 1977; 14: 2: 75-96.

66. Bock RD, Wood R. Test Theory. In: Annual Review of Psychology. PH Mussen and ME Rosenzweig (Eds.) 1971; 22: 193-224.

67. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, Mass.:Addison-Wesley, 1968.

68. Lord FM. In: Application of item response theory to practical testing problems; Lawrence Erlbaum Arson Hillsdale, N.J;1980.

69. Wright BD. Sample-free test calibration and person measurement. In: Proceedings of the 1967 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service;1968.

70. Swaminathan H. Parameter estimation in item response models. In: Applications of item response theory. RK Hambleton (Eds.);1983.

71. Torgerson WS. Theory and methods of scaling. New York: Wiley;1958.

72. McDonald RP. Factor analysis and related methods. Hillsdale, New Jersey: Lawrence Erbaum Associates;1985.

73. McDonald RP. The dimensionality of test and items. The British Journal of Mathematical and Statistical Psychology 1981;34:100-117.

74. Schiffman SS, Reynolds ML, Young FW. In: Introduction to multidimensional scaling. New York: Academic Press;1981.

75. Hattie J. An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research 1984; 19: 49-78.

76. Hulin CL, Drasgow F and Parsons CK. Item Response Theory: Application to Psychological Measurement. Homewood, Illinois: Dow Jones-Irwin 1983.

77. Stout WF. A new response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika 1990; 52: 589-617.

78. Nandakumar R. Traditional vs. essential dimensionality. J of Educational Measurement 1991; 28: 9-117.

79. Nanadakumar R, Stout W. Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics 1993; 18(1): 41-68.

80. DIMTEST Manual. Stout W, Douglas J, Junker B, Roussos L. A Fortran program for assessing dimensionality of binary item responses. Department of Statistics. University of Illinois at Urbana-Champaign April 1993.

81. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Jaeger RM, Eds.: Sage publications 1991.

82. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Statistical theories of mental test scores. FM Lord and MR Novick. Reading, Mass.:

Addison-Wesley 1968.

83. BILOG, Scientific Software, Inc. Illinois, Chicago, USA; 2nd Ed.1990.

84. Anderson EB. The numerical solution of a set of conditional estimation equations. The journal of the Royal Statistical Society 1972(b); 34: 42-54.

85. Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrica 1972; 37: 29-51.

86. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph 1969; No. 17.

87. Lord FM. Estimation of latent ability and item parameters when there are omitted responses. Psychometrica 1974(b); 39: 247-264.

88. Hambleton RK, Murray L. Some goodness of fit investigations for item response models. In: Applications of Item Response Theory. Educational Research Institute of British Columbia 1983; 71-94.

89. Hambleton RK and Swaminathan H. Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff Publishers 1983.

90. Bejar II. A procedure of investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement 1980; 17: 283-296.

91. Crocker L., Algina J. Item Analysis in Test Development. In: Introduction to classical and modern test theory. CBS College Publishing 1986.

92. Hulley SB, Cummings SR. Choosing the study subjects. In: Designing Clinical Research Methods: An Epidemiological Approach. Williams & Wilkins1992; 24-26.

Appendix 1
Measurement Terms

Construct: This term signifies an underlying trait or dimension of interest, and conceptually refers to an unobservable or latent, abstract variable (rather than a concrete one), e.g. anxiety, development. The domain of a construct is represented by items that may be its direct or indirect measures; the included items should provide an adequate representation of all aspects of the specified domain under study.

Validity: This refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. Validation procedure can be broken into the following three components:

Substantive Validation: This refers to the relevance of the components (newborn morbidity items) constituting the tool to the specified domain under examination e.g. relevance of each *newborn morbidity attribute* to the concept of *morbidity in newborns at birth.*

Structural Validation: This reflects a hypothesis that the components that make up the tool or outcome measure *correlate with one-another in studies of individual differences and have been shown to be similarly affected by experimental manipulations.*

External Validation: This refers to the *functional relations between the outcome measure and various behaviors that are external to the measure itself.* Determination of external validation with events from different fields supports utilization of the outcome measure in those fields.

Internal Consistency: This refers to the *homogeneity of components or items* that make-up the tool. It essentially reflects that all the included items in the tool are measuring the same trait(s).

Appendix 2
Assessment of Newborn Morbidity Items by Physicians

Title of the Study: 'Development of a Tool of Newborn Morbidity Index'

Physician Contact Phone No. Date

Do you wish to return the form with weighted items after the meeting?
Yes/No

Purpose:

Please find enclosed a list of items of morbidity in newborns. We want your opinion regarding the degree of morbidity at birth in terms of individual attributes of newborn morbidity that are provided in the adjoining list.

Exercise:

You are requested to grade individual items of newborn morbidity from the morbidity item-list, in which attributes of morbidity are categorized by the degree of the severity of morbidity.

The worst-physiology in the first 24 hours will be scored, unless otherwise stated along with items. Your score will reflect the degree of morbidity due to each morbidity item.

Appendix 2 (Contd.)

Example:

1. In your opinion, is it 2 or 2.5 times worse to have a mean B.P. of <30 m m of Hg for a newborn between 28-32 weeks of gestation as compared to having a B.P. of >30 mm of Hg for the same newborn? Say, you agree on 2.5. Record 2.5 in the item- list in the box corresponding to that item.

2. In your opinion, what is worse: to have a single seizure or multiple seizures? And how much worse (in comparative terms)? What, if it was refractory to 2 drugs? Make these comparisons to a situation of 'no seizures' being graded as '0'. Record your assessment in the adjoining form of newborn morbidity items.

Appendix 3
Physician's Assessment

Physician

Newborn Placement

Newborn Hospital ID

D.O.B.

Newborn Name

Newborn's Health Status	**Day 1 / Day 3 / Day 7 /Day of Discharge** (Please Circle Each Time)

Grade on the Analog Scale of Severity of Illness
(Please Circle Number 0-5)

DAY 1

No morbidity	Very Mild	Mild	Moderate	Severe	Most Severe
0	1	2	3	4	5

DAY 3

No morbidity	Very Mild	Mild	Moderate	Severe	Most Severe
0	1	2	3	4	5

DAY 7

No morbidity	Very Mild	Mild	Moderate	Severe	Most Severe
0	1	2	3	4	5

DAY OF DISCHARGE

No morbidity	Very Mild	Mild	Moderate	Severe	Most Severe
0	1	2	3	4	5

University of Alberta
Edmonton

Canada T5H 3V9

Department of
Obstetrics and Gynaecology

201 Community Services Centre
10240 Kingsway Avenue
Fax: (403) 477-4981

Appendix 4

NOTICE OF INFORMATION

To: **Neonatologists and Pediatricians attending NICU/ Regular Nursery Newborns, RAH and University of Alberta Hospital**

From: **Department of Obs & Gyne, RAH**

Phone: **477-4812 (Dr. Nan Okun)**

RE: **A Study of the tool of 'Newborn Morbidity Index'**

Background Information:

With the current low incidence of perinatal mortality in developed countries, significant morbidity at birth has become a more appropriate measure of perinatal outcome. Most obstetrical interventions produce subtle differences in individual neonatal morbid outcomes. Single outcome measures individually fail to reflect the overall newborn health status. As well, huge sample sizes are required to detect small differences in outcome when measures of serious but rare neonatal morbidity are examined. It is felt that measurement strategies need to catch up with improvements in outcomes in the last two decades.

There are no existing standardized measurement tools consisting of items of clinical attributes relevant to obstetrical interventions and therapeutic strategies. There is a need for a sensitive, valid and reliable measurement tool comprised of a scoring system of morbidities in neonates at birth, to serve as a basis for comparing modes of obstetrical therapies and interventions.

Salient Features:

1. The purpose of the tool is to be a new assessment index to quantify cumulative morbidity at birth in newborns.

2. A total of 600 newborns will be recruited from both NICU and regular nursery for the morbidity index data collection by research-assistants, starting November 95.

3. Charts of recruited newborns will be flagged and a sheet of the 'Analogue Scale' (see enclosed) will be attached.

4. We are requesting attending physicians to **mark their assessment** on the analogue scale, simultaneously with the initial newborn exam on day 1 and then on day 3, day 7, if the newborn continues to be under your care, and again on the day of discharge. Grading on the analogue scale should reflect "physician's assessment" of the overall morbidity status in terms of 'Very Mild', 'Mild', 'Moderate' and 'Severe' for that newborn.

5. Categorical and Precise information on the **Newborn's Tone** (Hypotonia Yes/No and when, if yes), **Apnea Status** and **Level of Consciousness** (Hyperalert/Drowsy/Lethargic OR Stupor/Obtundation OR Coma) will be extracted from your notes in the newborn's chart.

6. We will update you about the progress of the study at regular intervals. In case of any further clarification, please call us at the phone numbers noted above.

Appendix 5

University of Alberta
Edmonton

Canada T6G 2R7

Office of the Dean
Faculty of Medicine

2J2.00 WC Mackenzie Health Sciences Centre
Telephone: (403) 492-6621
Fax: (403) 492-7303
E-mail: _____@dean.med.ualberta.ca

RESEARCH ETHICS BOARD

ETHICS APPROVAL FORM

Dean
Dr D L I Tyrrell
492-9728
E-mail: ltyrrell@_____

Executive Assistant
Ms R M Spencer
492-1558
E-mail: rita@_____

Administrative Officer
Mr D M Whidden
492-6625
E-mail: don@_____

Associate Dean
Faculty Affairs
Dr R L Collins-Nakai
492-9727
E-mail: ruth@_____

Office of Research
Associate Dean
Dr J H Weiner
492-9723
E-mail: joel@_____

Assistant Dean
Dr W A McBlain
492-9720
E-mail: bill@_____

Postgraduate Medical Education
Associate Dean
Dr G Goldsand
492-9722
E-mail: ggldsnd@_____

Office of Admissions & Undergraduate Medical Education
2-45 Medical Sciences Building T6G 2H7
Phone: 492-6350
Fax: 492-9531

Associate Dean
Undergraduate Medical Education
Dr A H Wain
492-9523
E-mail: andi@_____

Assistant Dean
Clinical Education
Dr P M Crockford
492-9522

Assistant Dean
Admissions
Dr C I Cheeseman
492-9524
E-mail: cheese@_____

Date: June, 1996

Name(s) of Principal Investigator(s): Dr. Nanette Okun

Department: Obstetrics and Gynaecology

Title: Tool of Composite Newborn Morbidity Index in Term Newborns

The Research Ethics Board has reviewed the protocol involved in this project which has been found to be acceptable within the limitations of human experimentation. The REB has also reviewed and approved the patient information materials and consent form.

Specific Comments:

Signed - Chairman of Research Ethics Board

for the Faculty of Medicine
University of Alberta

This approval is valid for one year.

Issue #2002

Appendix 6
Guidelines and Operational Definitions for Data Collection

Inclusion Criteria:

1. Gestational age ≥28 weeks

2. For transfer admissions from the other hospitals to RAH - ensure availability of complete maternal antenatal and delivery record.

Record information on the following newborn morbidity items in the 'Data Collection Form for Newborn Morbidity Items' (Appendix 7):

A. Cord Blood pH
Assume normal if not done
- Enter as continuous as well as categorical variable
- From the Delivery Record OR Respiratory Laboratory in the NICU
- Refer 'Technical bulletin' of ACOG: November 1995: No. 216

B. Resuscitation at Birth
- Categorize
- From Delivery Record

C. Meconium-Cords
- Categorize
- From Delivery or Neonatal Record Sheet

D. and E. Apgar Score
- Categorize
- From Delivery Record

F. Heart Rate (within 24 hr. after birth)
- Consider the worst H.R. reading noted ≥ 2 times consecutively
- Categorize
- From the Neonatal Flow-Sheet

G. Systolic Blood Pressure (within 24 hr. after birth)
- Consider the worst systolic BP reading noted ≥ 2 times consecutively
- Mark category by gestational age

H. Respiratory Rate (within 24 hr. after birth)
- Consider the worst RR reading noted ≥ 2 times consecutively
- From the Neonatal Flow-Sheet
- Categorize

I. Color
- From the Neonatal Flow-Sheet

J. Hypotonia
• From the Neonatal Chart and Neonatal Flow-Sheet

K. Flaccidity
• Between 1-120 hr. after birth

L. Apneic Spells
(Periods of apnea of more than 20 seconds and/or those accompanied by bradycardia and cyanosis are termed apneic spells)
• ≥ 2 consecutive readings- From the Neonatal Flow-Sheet
• Categorize

M. Bleeding Disorder
• From the Neonatal Chart
• Any time before discharge

N. Seizures
• Categorize
• Any time before discharge
• From the Neonatal Chart
• Refer Neonatal-Perinatal Medicine: Fanaroff and Martin: 729-735; 1992)

O. Level of Consciousness (not iatrogenic)
• Categorize
• Any time before discharge
• From the Neonatal Chart and Flow-Sheet

P. Cardiopulmonary Resuscitation
• Any time before discharge
• From the Delivery Record and Neonatal Chart

Q. Oral Feeding Difficulties
• Categorize
• From the Neonatal Chart and Flow-Sheet

R. Respiratory Status
• Categorize
• From the Neonatal Chart and Flow-Sheet

S. Urine Output (any time before discharge)
• Categorize
• Low Urine Output is < 2 ml/kg/hour
• From the Neonatal Chart and Flow-Sheet

T. Trauma
• Categorize

• From the Delivery Record and Neonatal Chart

U. Hypoglycemia
• Assume normal, if not ordered
• Categorize - ≥2 readings at <1.7 mmol/L or <2.2 mmol/L (with dextrostix)
• From the Neonatal Flow-Sheet

V. Hyperbilirubinemia
• Record as categorical variable
• From the Laboratory Test Reports

W. Bacterial Culture
• From the Laboratory Test Reports
• Assume negative, if not ordered

X. Intra-Ventricular Hemorrhage
• From the Physician's notes and discharge summary

Maternal Events

Record information on the following maternal events from the Pre-natal Chart or Delivery Records in the 'Data Collection Form for Maternal Events' (Appendix 8):

a) Diagnosis of fetal distress (defined as a prolonged series of late or variable decelerations and a decrease or loss of fetal heart rate variability (<2 accelerations in 20 minutes of ≥15 bpm) on electronic fetal heart monitoring in the intrapartum period)

b) Pre-gestational diabetes

c) Gestational diabetes mellitus - (i) controlled by diet, (ii) controlled by diet and insulin

d) Preeclampsia - diagnosed by the attending physician at a BP of >140/90 mm of Hg on >2 occasions within 30 minutes or >30/15 mm of Hg change and proteinuria

e) Hemolysis, Elevated Liver Enzyme and Low Platelet (HELLP) syndrome

Chorioamnionitis (clinical diagnosis)

f) Maternal Obesity: Pre-pregnancy Weight >90kg

g) Breech Delivery

h) Maternal Chronic Disease: i. renal ii. connective tissue disease iii. epilepsy

i) Maternal Infections: i. Chorioamnionitis ii. GBS positivity

j) Antepartum Hemorrhage

k) Oligohydramnios/Polyhydramnios

Study No._____

Appendix 7
Newborn Data Collection Form

Dates Form Scored		Investigator	
Name		Mother's Age	
ID			
Gender		Gestational Age & Parity	
Physician			
		Date & Time of Birth	
D.O.A.	A Side B Side	C Side	D.O.D.
Home Address			
& Phone			
		day 1 day 3 day 7	discharge
Remarks		Discharge Status	Alive/dead

SNo	Item	Check	SNo	Item	Check
A	Cord Blood pH		D	Apgar Score (5 minute)	
	>7.10	0		7-10	0
a	7.0-7.1		*a*	4-6	
b	<7.0		*b*	1-3	
B	Resuscitation at birth		*c*	0	
	None	0	E	Apgar Score (10 minute)	
a	Bag & mask			7-10	0
b	Intubation		*a*	4-6	
C	Meconium		*b*	1-3	
	None	0	*c*	0	
a	Meconium above cords				
b	Meconium below cords				

Study No._____

Appendix 7 (Contd.)
Newborn Data Collection Form

SNo	Item	Check
F	**Heart Rate/minute**	
	100-160	0
a	161-200	
b	>200	
c	<100	
G	**Systolic B.P. (mean, mm of Hg)**	
	28-32 weeks 32-42 weeks	
	30-60 40-65	0
a	<30 <40	
H	**Respiratory Rate**	
	R.R./minute (30-60) for >2	0
	consecutive readings in first 3 hr.	
a	R.R./minute (<30 or >60)	
b	R.R.> (100/min) 3 -24 hr.	
c	R.R./minute [<30 or >60] in 3-24 hr.	
I	**Colour**	
a	Peripheral Cyanosis/Ruddy Colour	
b	Central Cyanosis/Dusky Colour	
J	**Hypotonia**	
	None (Normal)	0
a	Present (Identified) at <1 hr.	
b	Present between 1-120 hr.	
c	Present beyond 120 hr.	
K	**Flaccidity (between 1-120 hrs)**	
a	Present	
L	**Apnea**	
a	Apnea with spontaneous recovery	
b	pnea and Need for Oxygen therapy	
c	Apnea and Need for Resuscitation	

SNo	Item	Check
M	**Bleeding Disorder**	
a	Bleeding disorder G.I/Lungs/Skin	
b	Need for transfusion	
N	**Seizures**	
	None (normal)	0
a	Jittery/Tremors	
b	Non-Recurring Single Seizure (once)	
c	Multiple Seizures	
d	N.c and >2 drugs used for treatment	
O	**Level of consciousness**	
	Alert (normal)	0
a	Hyperalert/Hyper-irritable/Hypertonic	
b	Drowsy/Lethargic	
c	Stupor/Obtundation/Coma	
P	**Cardio-Pulmonary Resuscitation**	
a	Any time before discharge	
Q	**Oral Feeding Difficulties**	
	None (Normal)	0
a	Poor sucking within 24hr.	
b	Poor sucking between 24 hr.-7 days	
c	Poor sucking beyond day 7	
d	Persistent Vomitting	
R	**Respiratory Status**	
	Room Air	0
a	Need for assisted ventilation within 24 hr.	
b	Assisted ventilation after 24 hr.	
c	Mechanical ventilation within 24 hr.	
d	Mechanical ventilation in 24 hr.-7 days	
e	Mechanical ventilation beyond day 7	

Study No._____

Appendix 7 (Contd.)
Newborn Data Collection Form

SNo	Item	Check
S	**Urine Output (while in hospital)**	
	Normal	0
a	Abnormal (low)	
T	**Trauma**	
	None	0
a	Cephalhematoma	
b	Long Bone/Clavicle #	
c	Skull #	
	Facial/Peripheral Nerve Injury	
d	Not present at discharge	
e	Present at discharge, but improving	
f	Present at discharge, not improving	
g	Subdural Hematoma	
h	Intracerebral Hematoma	
i	Spinal Cord Injury	
List the following:		
Congenital Anomaly/Syndromes		
Inborn Error of Metabolism		

SNo	Item	Check
U	**Hypoglycemia (lowest level)**	
	None (Normal)	0
a	Blood Glucose (1.7-2.2 mmol/L)	
b	Definite Hypoglycemia(<1.7 mmol/L)	
V	**Hyperbilirubinemia (peak level)**	
	Bilirubin < 170 units (normal)	0
a	Bilirubin 170-250	
b	Bilirubin >250/Phototherapy	
c	Bilirubin >340/Exchange Transfusion	
W	**Bacterial Culture**	
	Negative	0
a	Blood Positive	
b	CSF Positive	
X	**Intra-ventricular Hemorrhage**	
	None	0
a	Grade 1 and 2	
b	Grade 3 and 4	
Total Morbidity Score		

Appendix 8
Data Collection Form for Maternal Events

Name		
Diagnosis (if any)		
Mode of Delivery	_	**Vaginal**
	_	**Cesarean Section**

a)	Fetal Distress	No/Yes; If yes, please describe
b)	Pre-gestational Diabetes Mellitus	No/Yes
c)	Gestational Diabetes Mellitus	No/Yes, if yes Grade i. Diet Therapy Grade ii. Diet+ Insulin Therapy
d)	Preeclampsia	No/Yes
e)	HELLP syndrome	No/Yes
f)	Maternal Obesity Pre-pregnancy Weight >90 kg	No/Yes
g)	Breech Delivery	No/Yes
h)	Maternal Chronic Disease	No/Yes i. renal ii. connective tissue disease iii. epilepsy
i)	Maternal Infections	No/Yes i. chorioamnionitis ii. GBS positivity
j)	Antepartum Hemorrhage	No/Yes; if yes, please describe
k)	Oligohydramnios/Polyhydramnios	No/Yes

Capital Health Authority
Public Health Services

Appendix 9
Healthy Beginnings Postpartum Program
Flow Sheet

✓	Normal/Not Significant
•	Observe or note
R	Referred
NA	Not Applicable
X	Not Addressed

INFANT NAME _____ D.O.B. _____
Surname First Year Month

ADDRESS _____ HOSPITAL _____

POSTAL CODE _____ GEO CODE _____ CHILD'S PHYSICIAN _____

TELEPHONE NO: _____ WILL ATTEND _____ HEALTH CE

MOTHER'S NAME: _____ FATHER'S NAME: _____
Surname First Surname Fi

SIBLINGS	DATE OF BIRTH	HEALTH PROBLEMS

SIBLINGS:	DATE OF BIRTH	HEALTH PROB

MOTHER'S HEALTH	1st	PAT
Prenatal Care/Classes		
Pregnancy		
Labour/Delivery		
Lochia		
Fundus		
Perineum		
Bowels		
C/S Incision		
Extremities		
Nutrition		
Sleep/Rest		
Coping/Parent		
Postpartum Exercise		
Family Planning		
Parental Immunization Status		
TPR		
Other		

FEEDING ASSESSMENT	1ST		PAT	
Breastfeeding ☐ Formula ☐	Observed		Observed	
	Yes	No	Yes	No
Breastfeeding				
Formula Feeding				
Parent's Concerns				
Comfort/Breast				
Latch/Position				
Frequency/Duration				
Urine Output				
Stool Output				
Supplements/Aids				
Weight/Growth				
Anticipatory Guidance				
Other				

INFANT'S HEALTH	1ST
Cord	
Circumcision	
Skin	
Sleep	
Injury Prevention (I.P.)	
Environment	
Resources Required	
Follow-up	
Immunization Sheets Given	
Hepatitis B Status	
Family History	
Other	

Date: _____ Signature of Nurse First Visit _____

Date: _____ Signature of Follow - Up Contact _____

Appendix 9 (contd.)

Nursing Assessment of the Newborn

Code

- ✓ Normal/Not Significant
- • Observe or note
- R Referral
- NA Not Applicable
- X Not Addressed

INFANT NAME _____ D.O.B. _____

Surname First Year Month D

GENERAL APPEARANCE	In	F/U?
Alert		
Smiling		
Symmetry		
Movements		
Other		
FACE	**In**	**F/U?**
Nose		
Mouth		
Other		
EYES	**In**	**F/U?**
Eye Lids		
Conjunctiva		
Other		
EARS	**In**	**F/U?**
Position		
Other		
HEAD	**In**	**F/U?**
Symmetry		
Shape		
Scalp		
Hair		
Fontanel A		
Fontanel P		
Neck		
Clavicles		
Other		

SKIN	In	F/U?
Colour		
Turgor		
Rashes		
Birthmark(s)		
Texture		
Nails		
Body Hair		
CHEST	**In**	**F/U?**
Symmetry		
Shape		
Breasts		
Nipples		
Apical Pulse		
Heart Sounds		
Respirations		
Other		
ABDOMEN	**In**	**F/U?**
Symmetry		
Shape	IST	F/U?
Spine		
Other		
BACK		
Spine		
Buttocks		
Other		

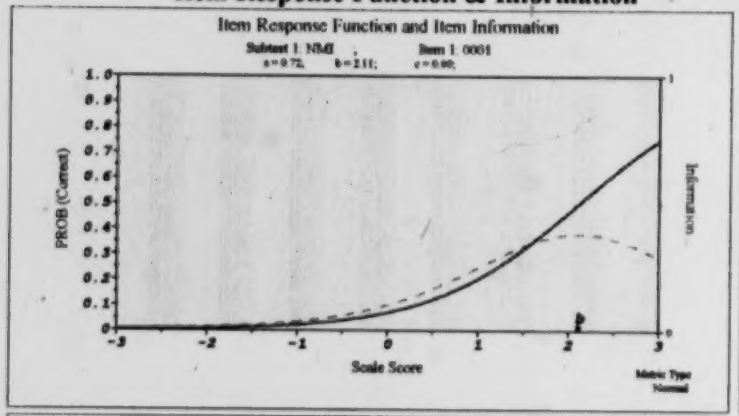
EXTREMITIES	In
Symmetry	
Mobility	
Muscle Tone	
Hips	
ANAL, GENITO-URINARY	**In**
Female - Appearance	
- Other	
Male - Appearance	
- Testes	
- Circumcision	
- Other	
HBPP ASSESSMENTS	**In**
Metabolic Screen	
Temperature	
Head Circumference	
Length	
Other	
PARENT/CHILD INTERACTION	**In**
Verbal Comments	
Eye Contact	
Holding	
Other	
SIX WEEK CHECK-UP	**In**
Mother	
Infant	

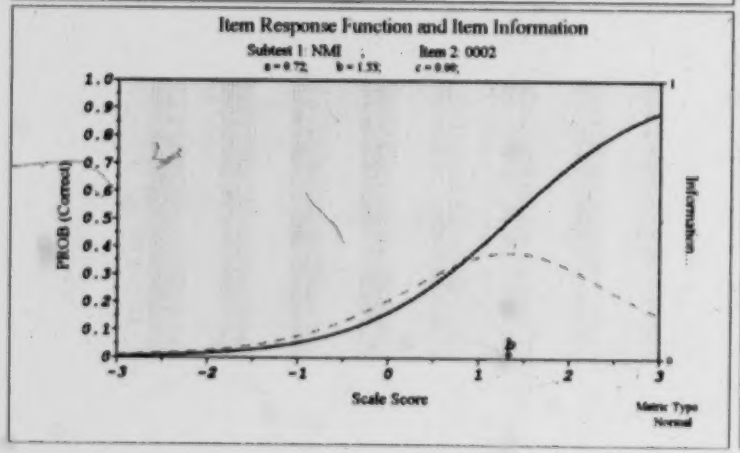
Date _____ Signature of Nurse First Visit _____

Date _____ Signature of Follow-Up Contact _____

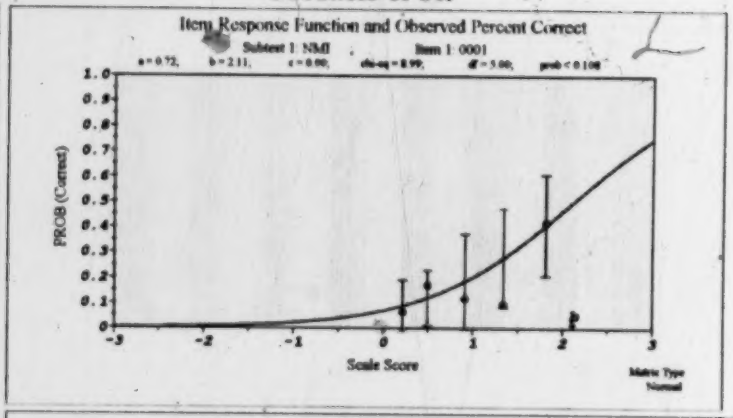
Appendix 10
Item Response Function Curves
Binary Item Numbers 1-50

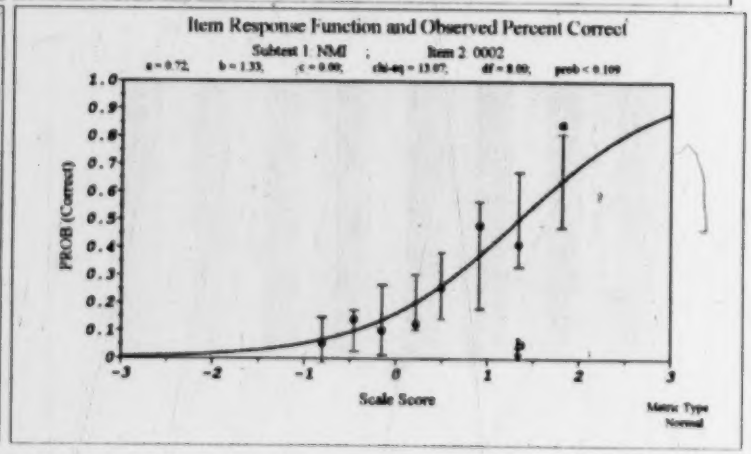
Item Response Function & Information





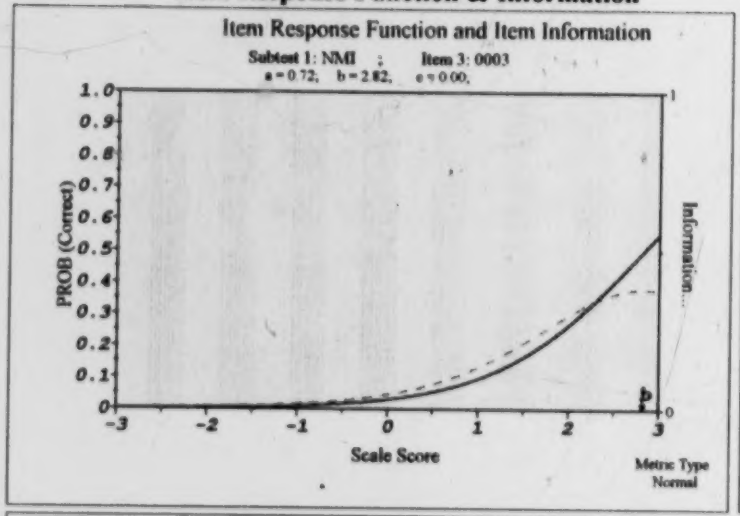
Goodness-of-Fit

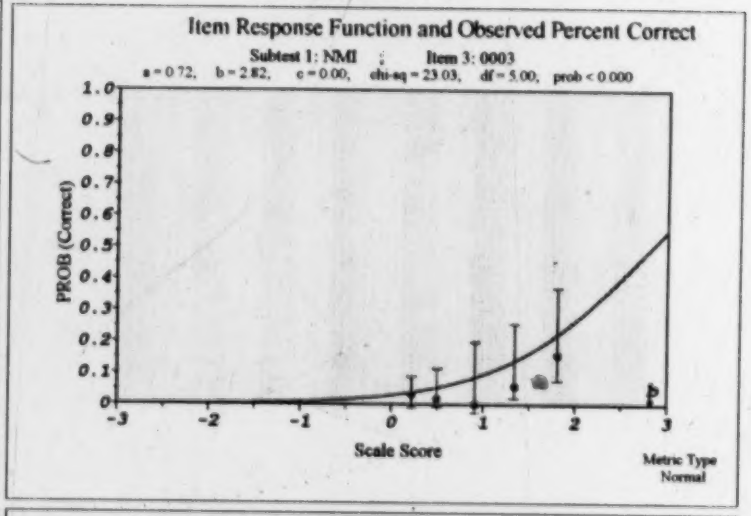

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 1: 0001
a = 0.72; b = 2.11; c = 0.00; chi-sq = 8.99; df = 5.00; prob < 0.108
PROB (Correct)
Scale Score
Metric Type
Normal

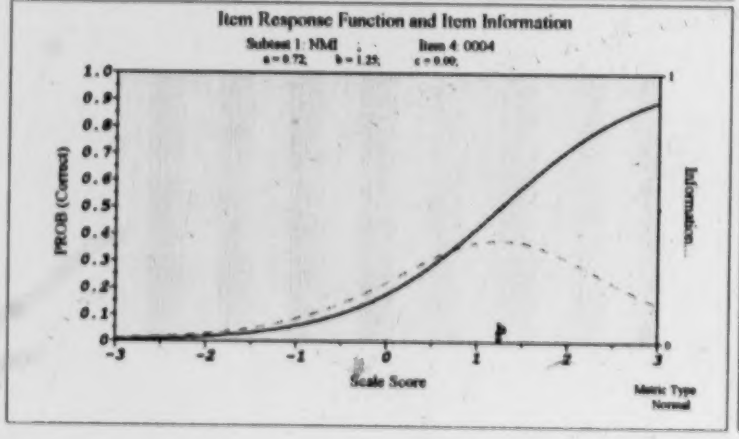

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 2: 0002
a = 0.72; b = 1.33; c = 0.00; chi-sq = 13.07; df = 8.00; prob < 0.109
PROB (Correct)
Scale Score
Metric Type
Normal

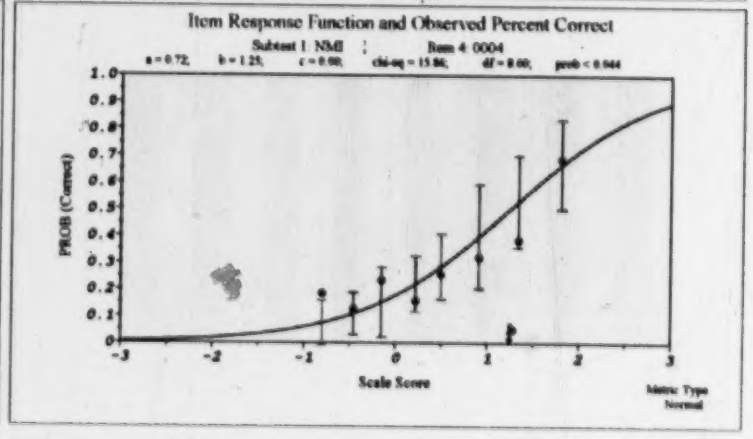
Item Response Function & Information

Goodness-of-Fit

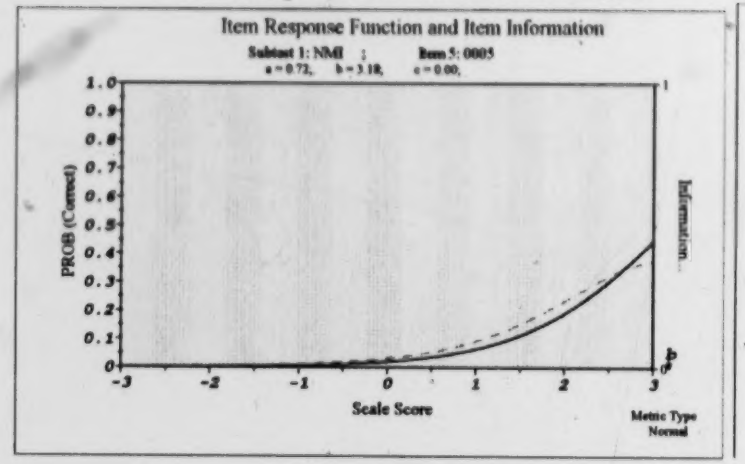



Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 3: 0003
a = 0.72; b = 2.82; c = 0.00; chi-sq = 23.03; df = 5.00; prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

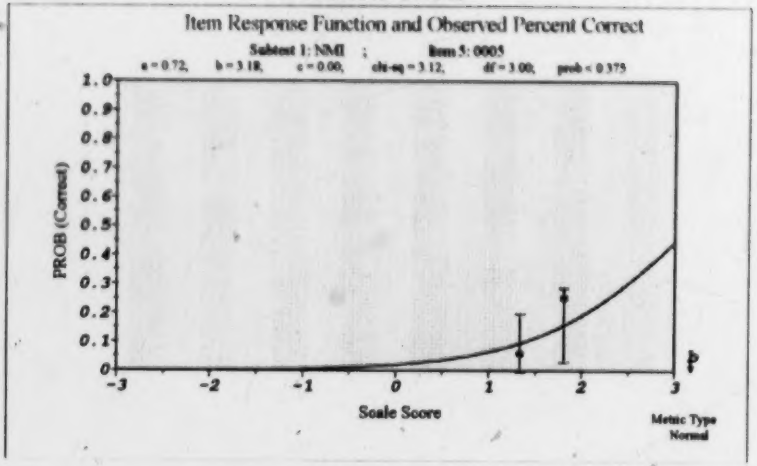


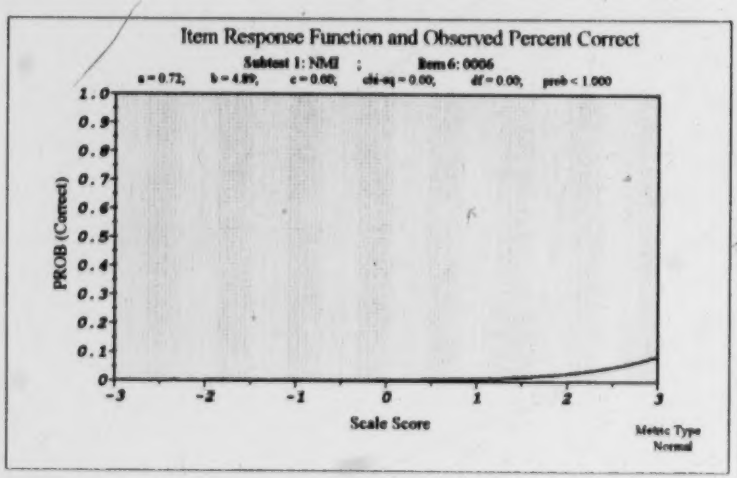

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 4: 0004
a = 0.72; b = 1.25; c = 0.00; chi-sq = 15.86; df = 8.00; prob < 0.044
PROB (Correct)
Scale Score
Metric Type
Normal

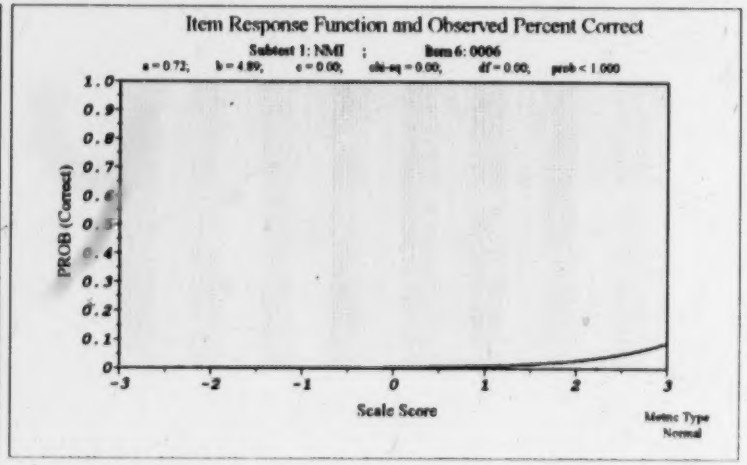
Item Response Function & Information



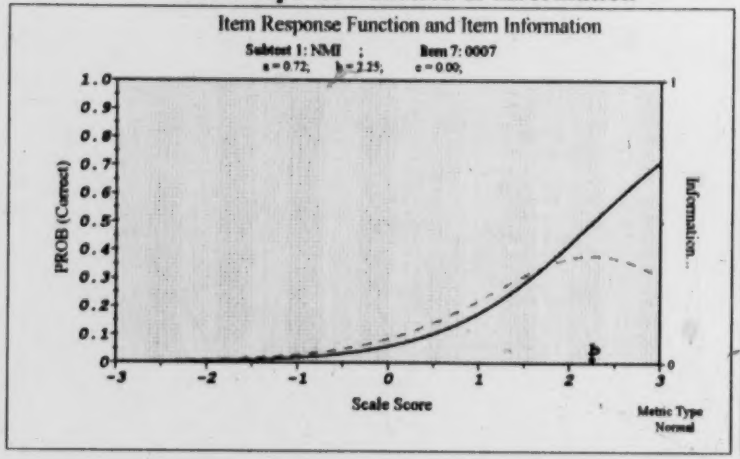
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 5: 0005
a = 0.72, b = 3.18, c = 0.00; chi-sq = 3.12, df = 3.00, prob = 0.375
PROB (Correct)
Scale Score
Metric Type
Normal

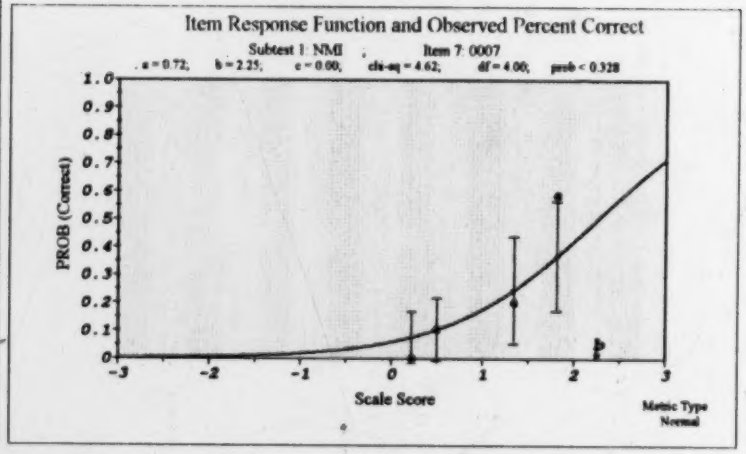

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 6: 0006
a = 0.72, b = 4.89, c = 0.00; chi-sq = 0.00, df = 0.00, prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

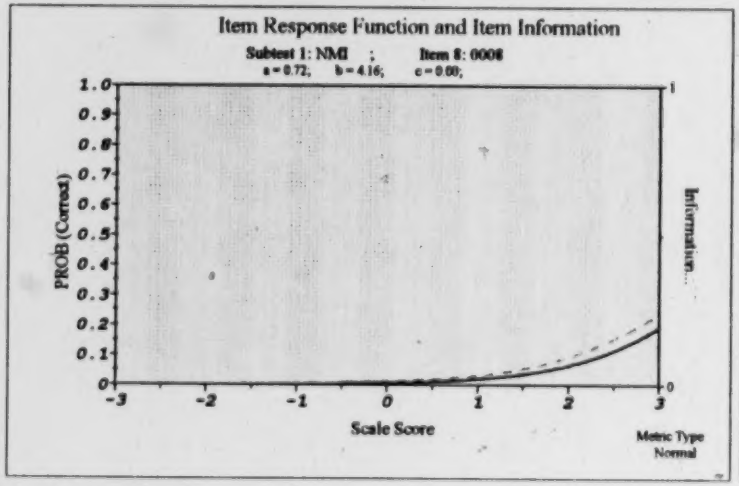

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 6: 0006
a = 0.72, b = 4.89, c = 0.00; chi-sq = 0.00, df = 0.00, prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

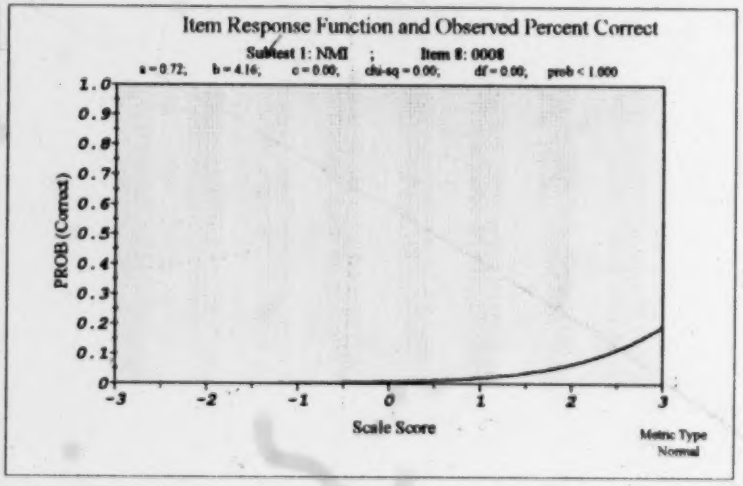
Item Response Function & Information



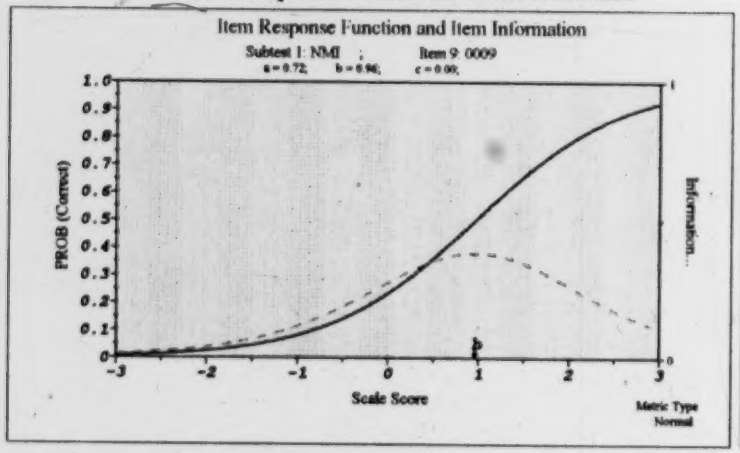
Goodness-of-Fit

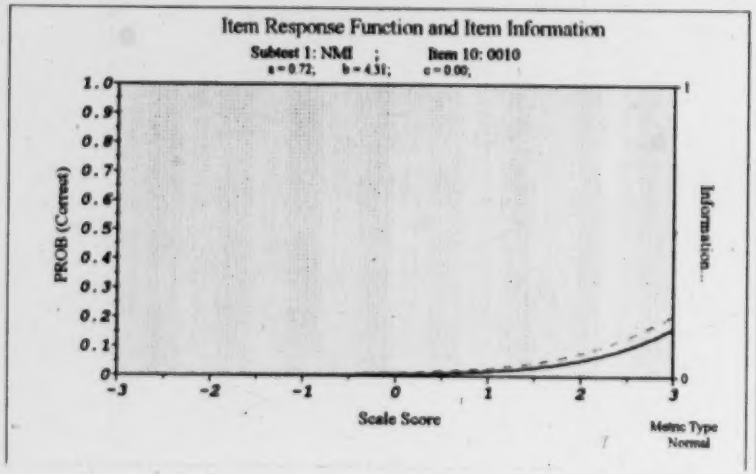

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 7: 0007
a = 0.72; b = 2.25; c = 0.00; chi-sq = 4.62; df = 4.00; prob < 0.328
PROB (Correct)
Scale Score
Metric Type
Normal



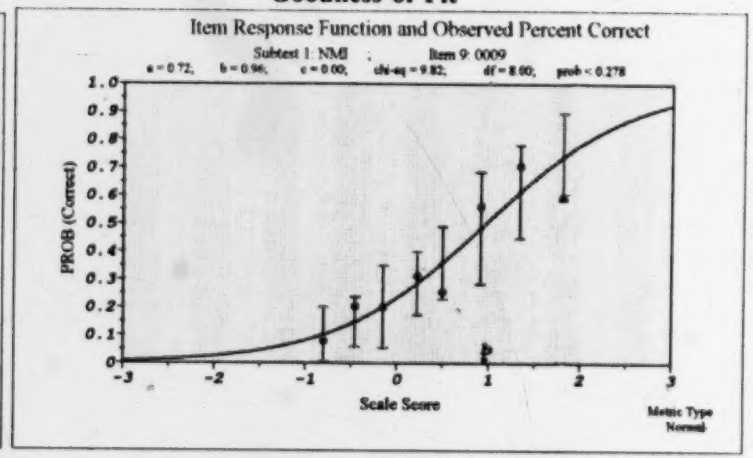

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 8: 0008
a = 0.72; b = 4.16; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

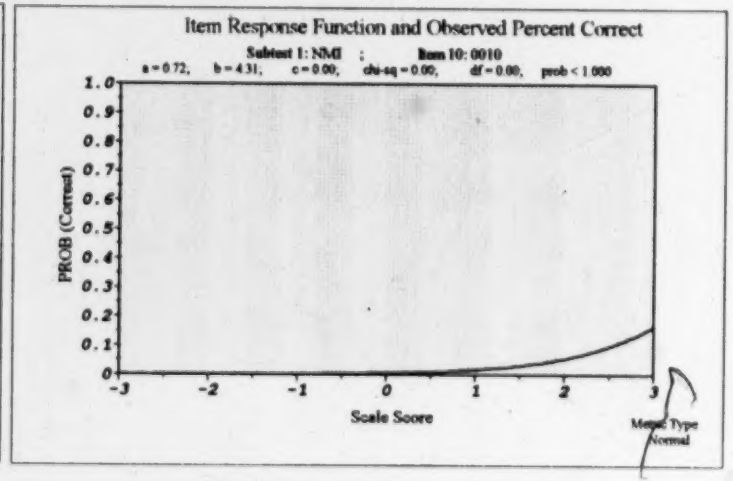
Item Response Function & Information



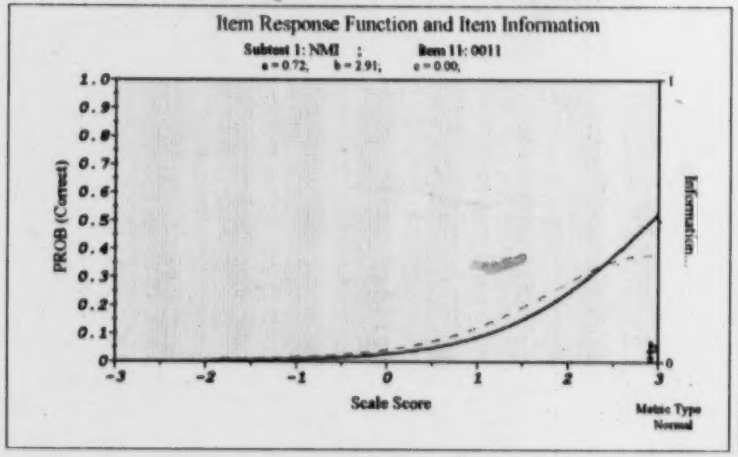


Goodness-of-Fit

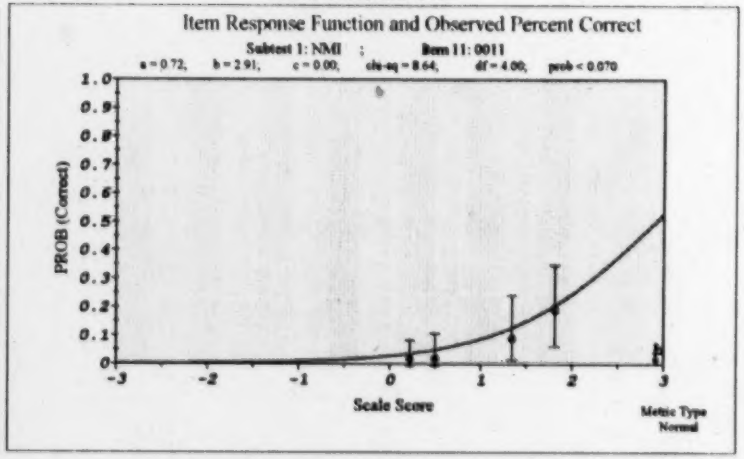

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 9: 0009
a = 0.72; b = 0.96; c = 0.00; chi-sq = 9.82; df = 8.00; prob < 0.278
PROB (Correct)
Scale Score
Metric Type
Normal

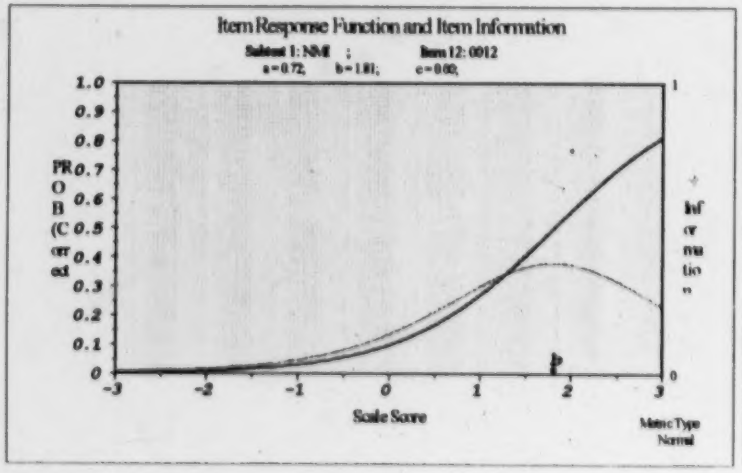

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 10: 0010
a = 0.72; b = 4.31; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

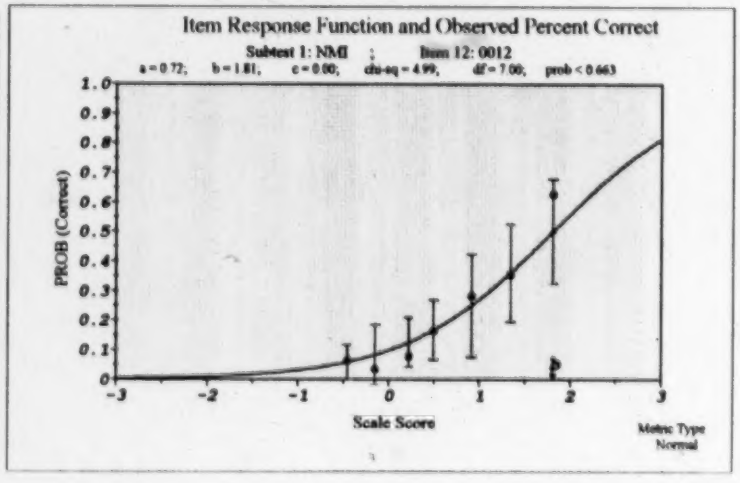
Item Response Function & Information



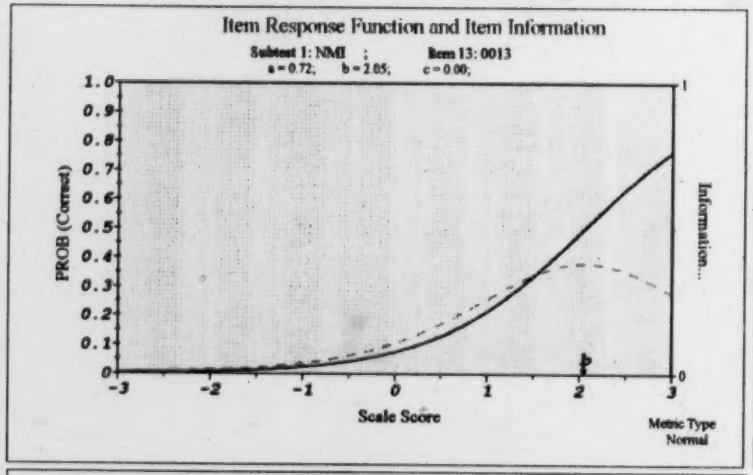
Goodness-of-Fit

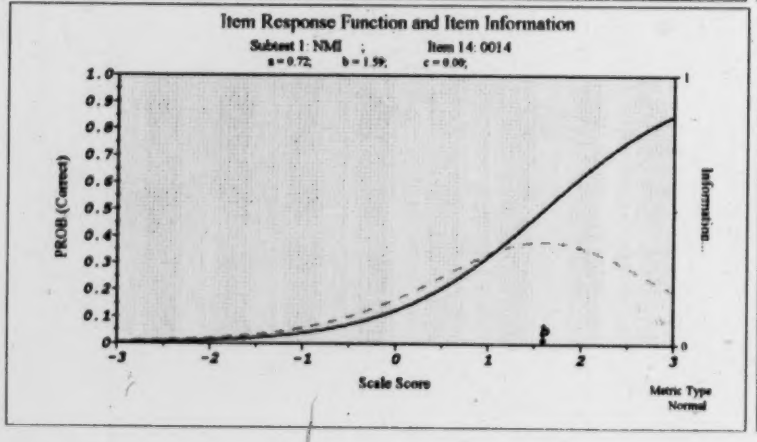

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 11: 0011
a = 0.72, b = 2.91, c = 0.00, chi-sq = 8.64, df = 4.00, prob < 0.070
PROB (Correct)
Scale Score
Metric Type
Normal



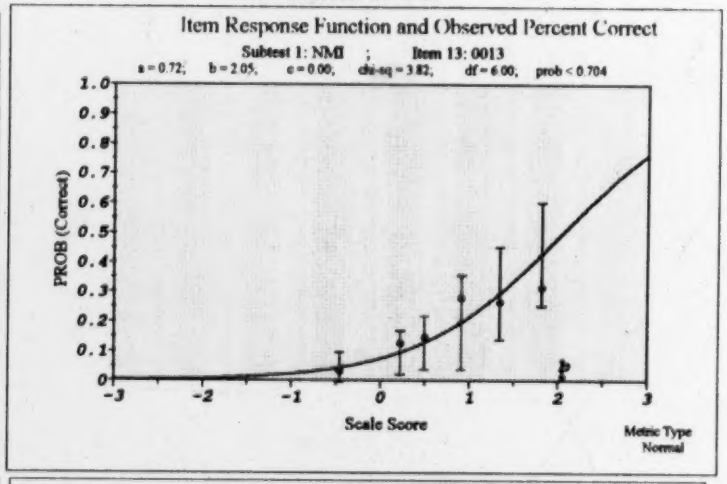

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 12: 0012
a = 0.72, b = 1.81, c = 0.00, chi-sq = 4.99, df = 7.00, prob < 0.663
PROB (Correct)
Scale Score
Metric Type
Normal

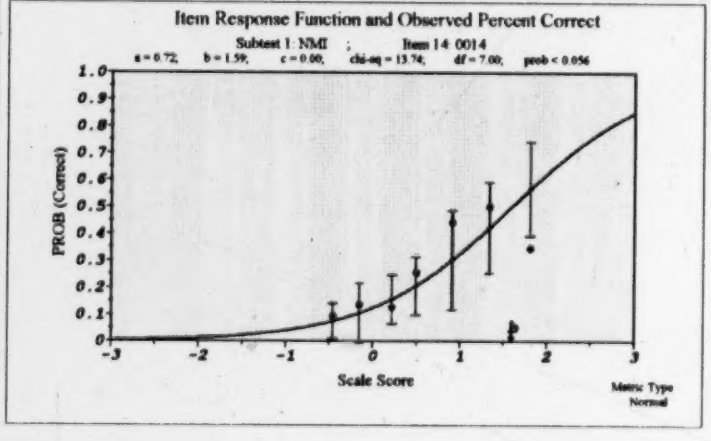
Item Response Function & Information



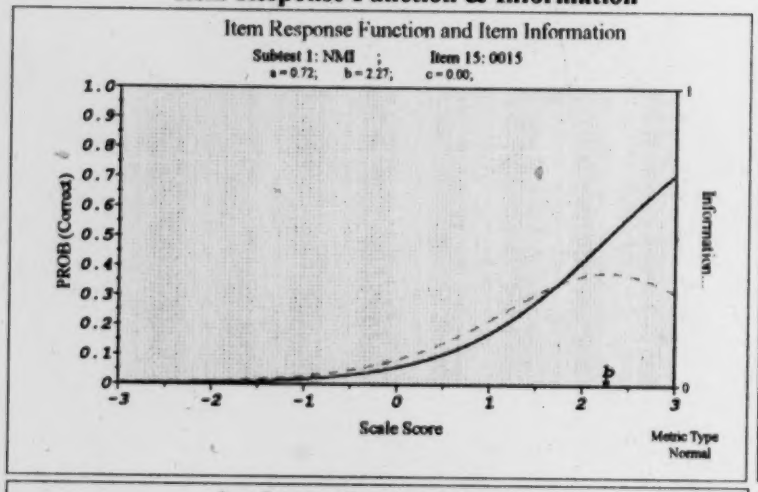


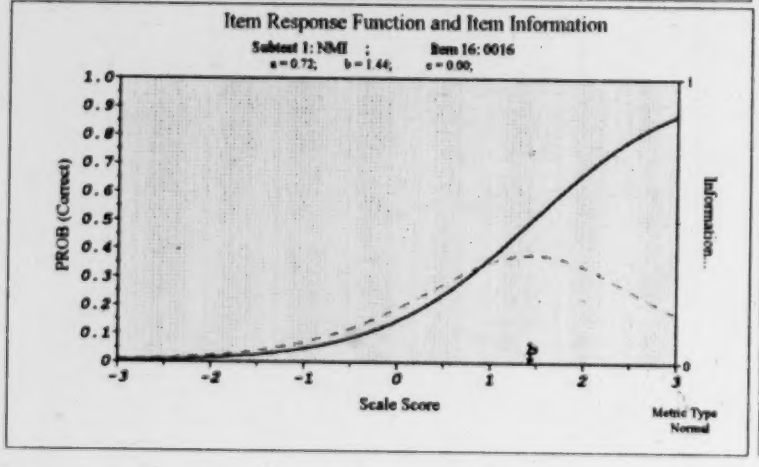
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 13: 0013
a = 0.72; b = 2.05; c = 0.00; chi-sq = 3.82; df = 6.00; prob < 0.704
PROB (Correct)
Scale Score
Metric Type
Normal

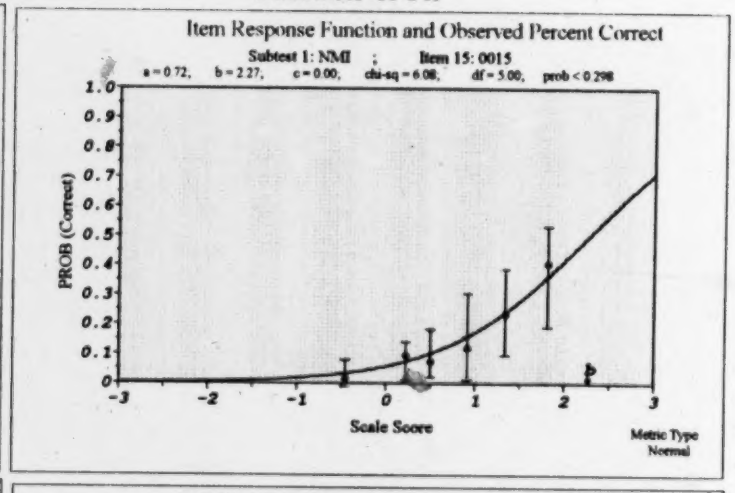

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 14: 0014
a = 0.72; b = 1.59; c = 0.00; chi-sq = 13.74; df = 7.00; prob < 0.056
PROB (Correct)
Scale Score
Metric Type
Normal

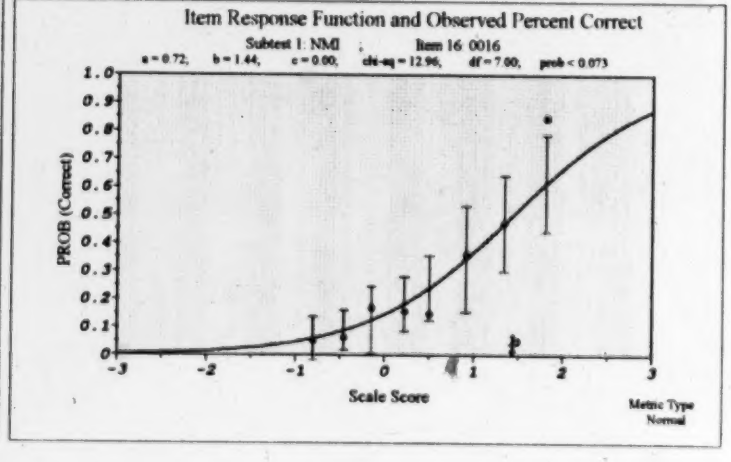
Item Response Function & Information



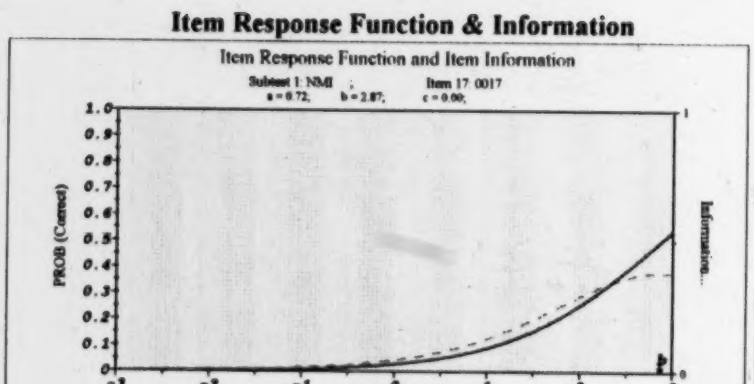


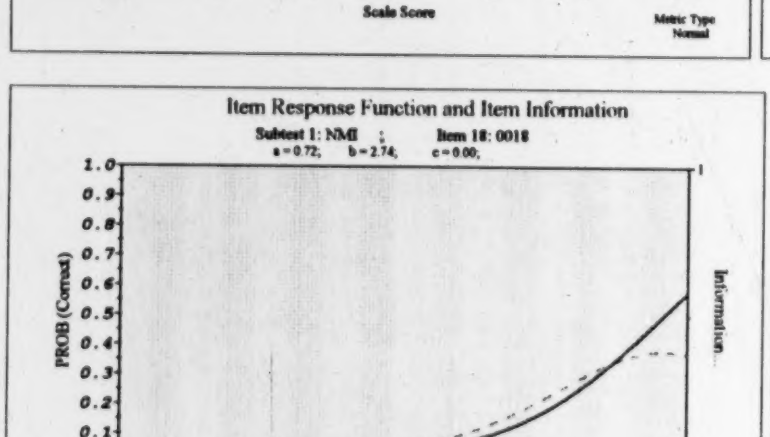
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 15: 0015
a = 0.72; b = 2.27; c = 0.00; chi-sq = 6.08; df = 5.00; prob < 0.298
PROB (Correct)
Scale Score
Metric Type
Normal

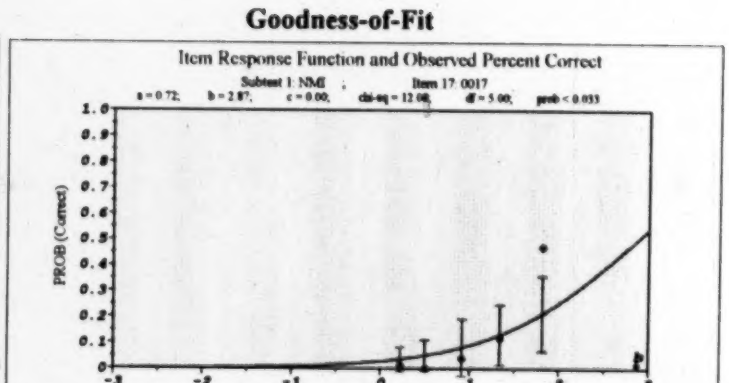

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 16: 0016
a = 0.72; b = 1.44; c = 0.00; chi-sq = 12.96; df = 7.00; prob < 0.073
PROB (Correct)
Scale Score
Metric Type
Normal

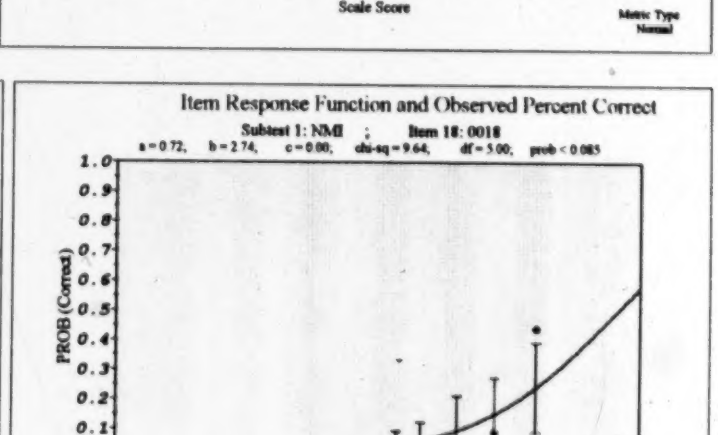
Item Response Function & Information



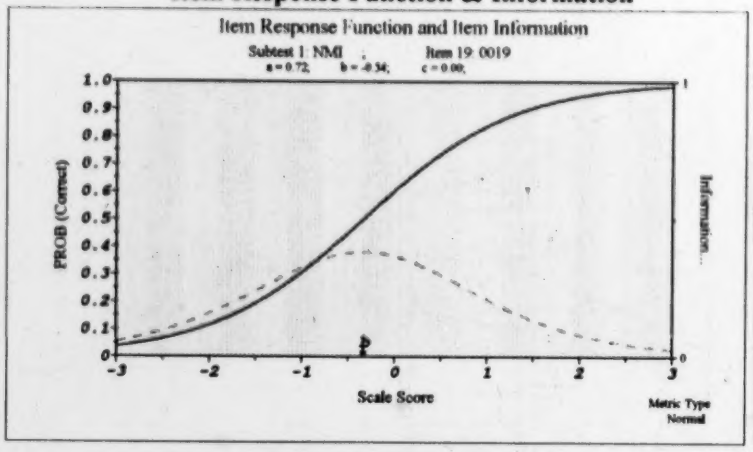


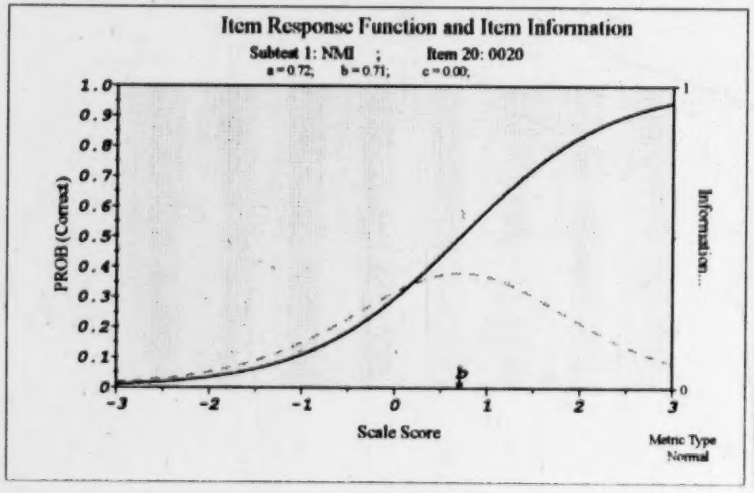
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 17: 0017
a = 0.72; b = 2.87; c = 0.00; chi-sq = 12.08; df = 5.00; prob < 0.033
PROB (Correct)
Scale Score
Metric Type
Normal

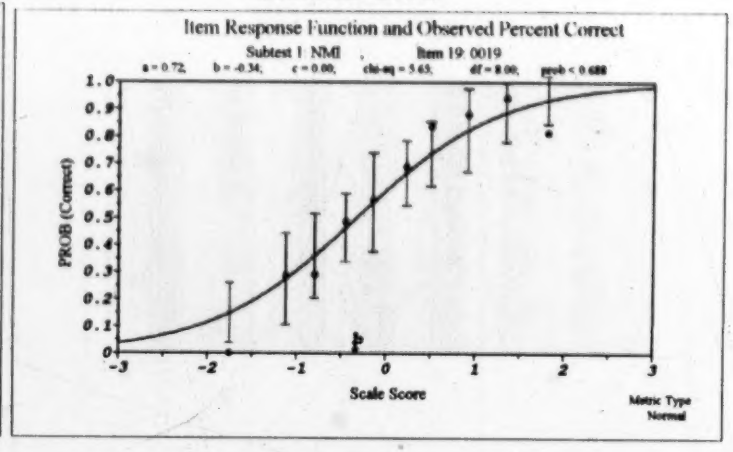

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 18: 0018
a = 0.72; b = 2.74; c = 0.00; chi-sq = 9.64; df = 5.00; prob < 0.085
PROB (Correct)
Scale Score
Metric Type
Normal

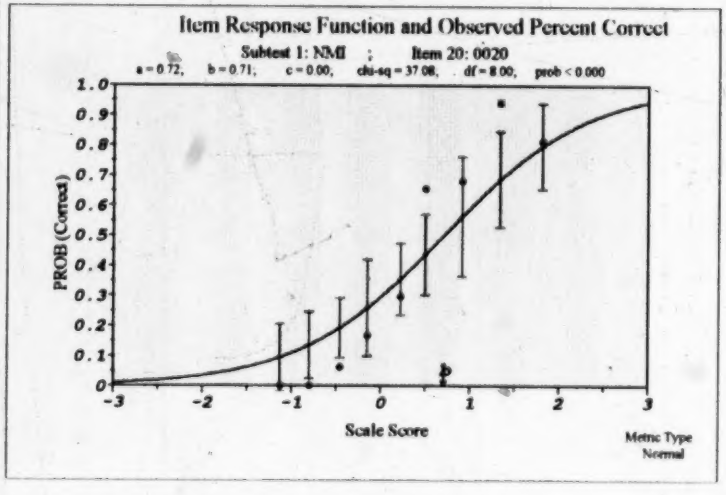
Item Response Function & Information



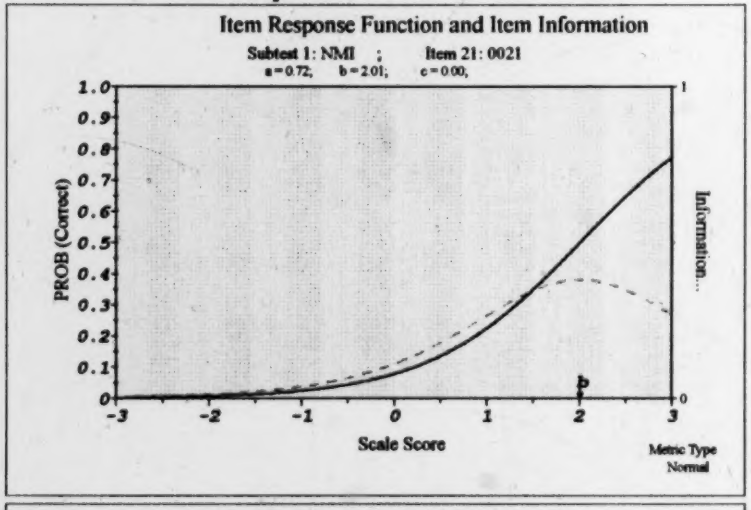


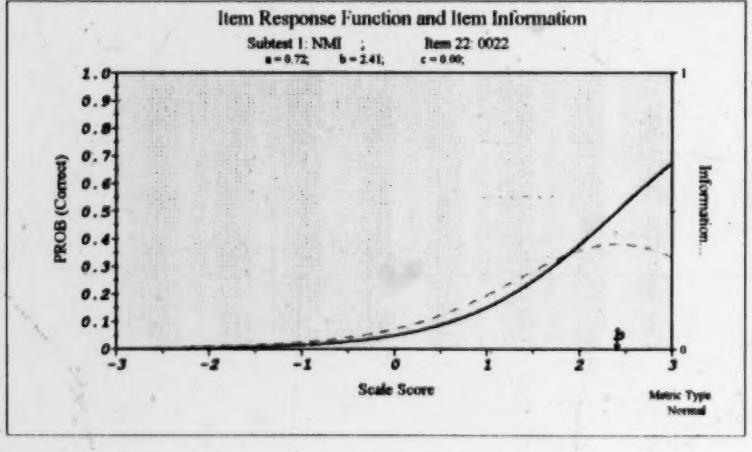
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 19: 0019
a = 0.72; b = -0.34; c = 0.00; chi-sq = 5.65; df = 8.00; prob < 0.688
PROB (Correct)
Scale Score
Metric Type
Normal

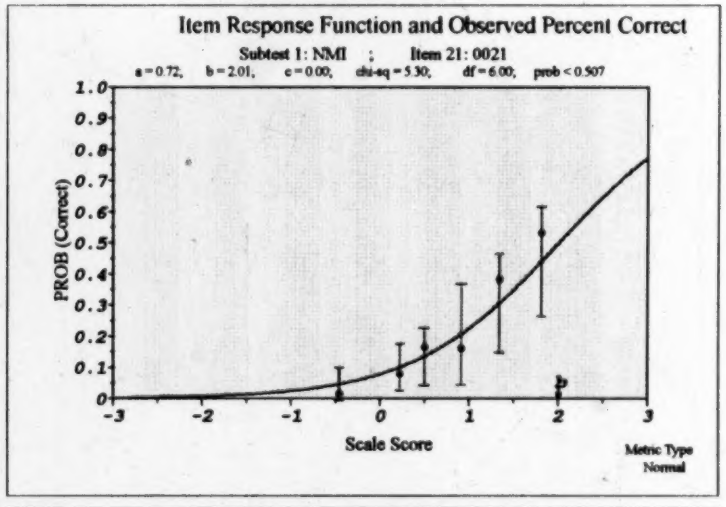

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 20: 0020
a = 0.72; b = 0.71; c = 0.00; chi-sq = 37.08; df = 8.00; prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

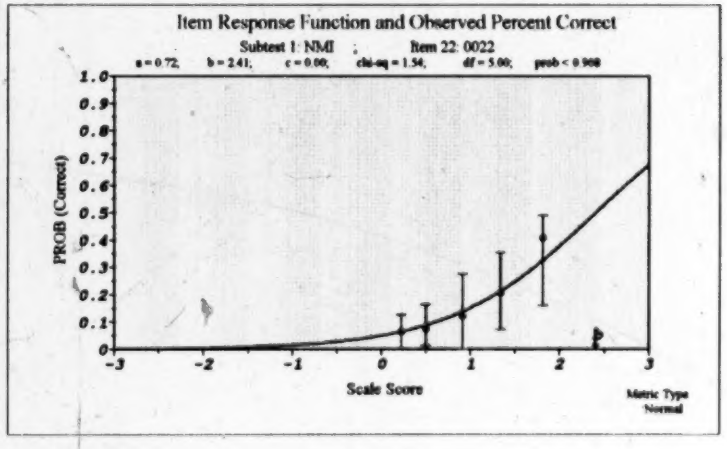
Item Response Function & Information



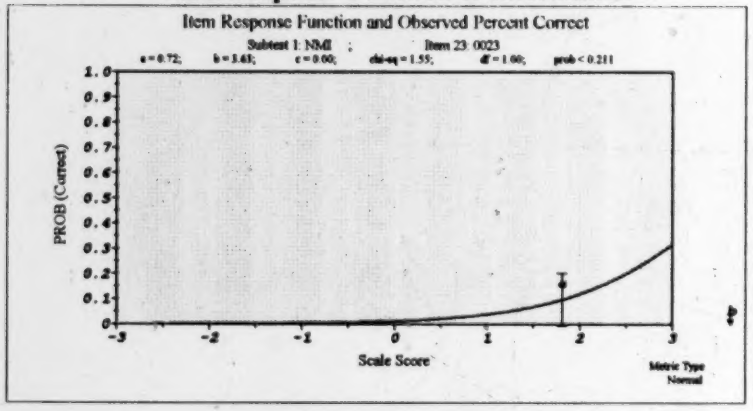


Goodness-of-Fit

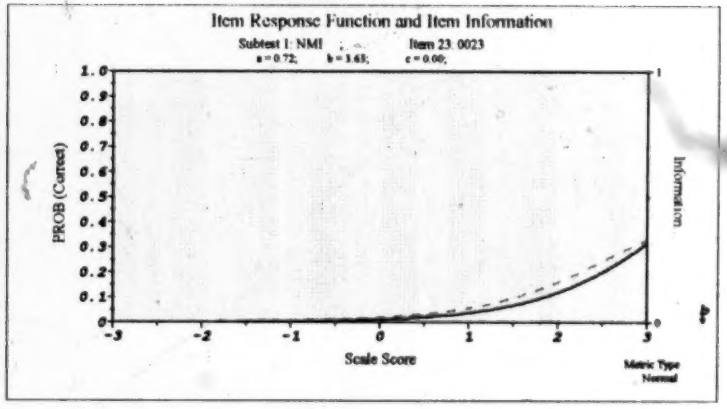

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 21: 0021
a = 0.72; b = 2.01; c = 0.00; chi-sq = 5.30; df = 6.00; prob < 0.507
PROB (Correct)
Scale Score
Metric Type
Normal

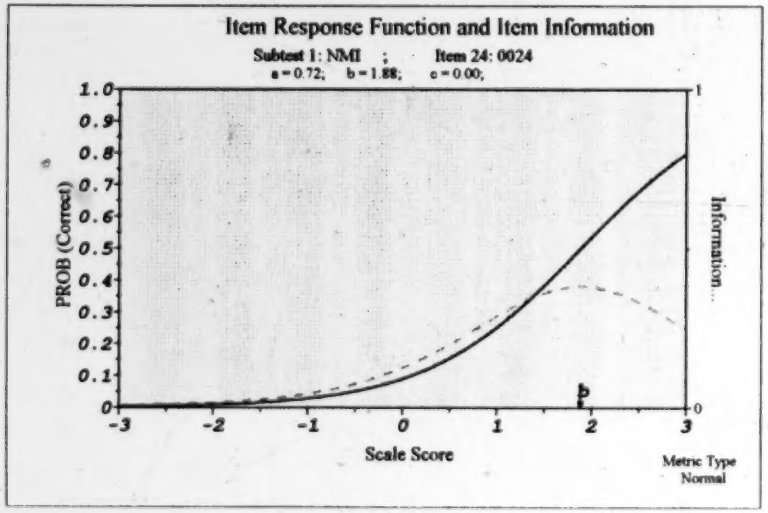

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 22: 0022
a = 0.72; b = 2.41; c = 0.00; chi-sq = 1.54; df = 5.00; prob < 0.908
PROB (Correct)
Scale Score
Metric Type
Normal

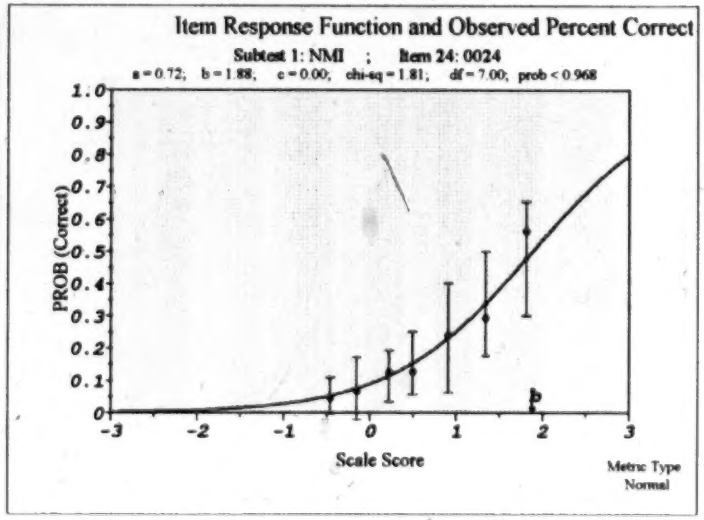
Item Response Function & Information


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 23: 0023
a = 0.72; b = 3.63; c = 0.00; chi-sq = 1.55; df = 1.00; prob < 0.211
PROB (Correct)
Scale Score
Metric Type
Normal

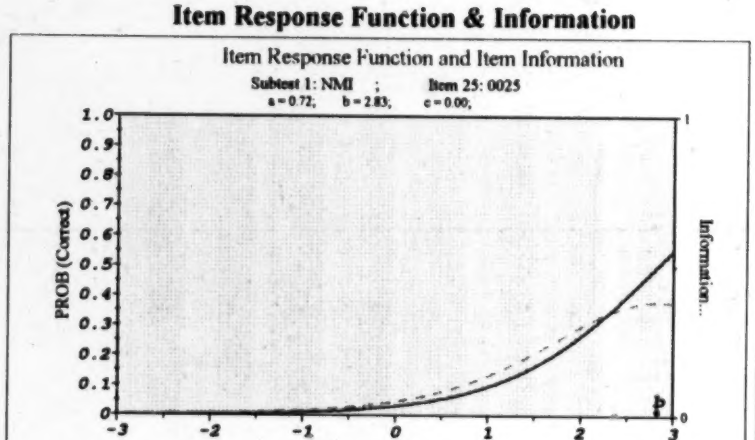
Goodness-of-Fit

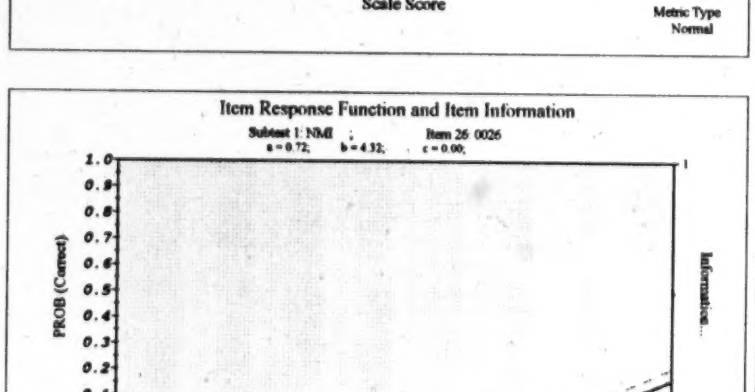




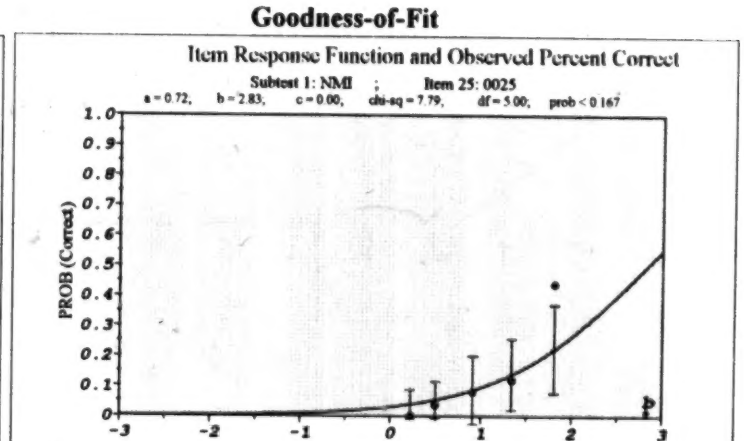

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 24: 0024
a = 0.72; b = 1.88; c = 0.00; chi-sq = 1.81; df = 7.00; prob < 0.968
PROB (Correct)
Scale Score
Metric Type
Normal

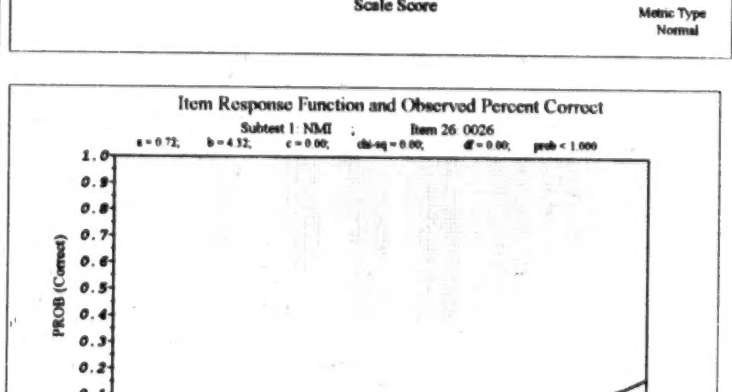
Item Response Function & Information



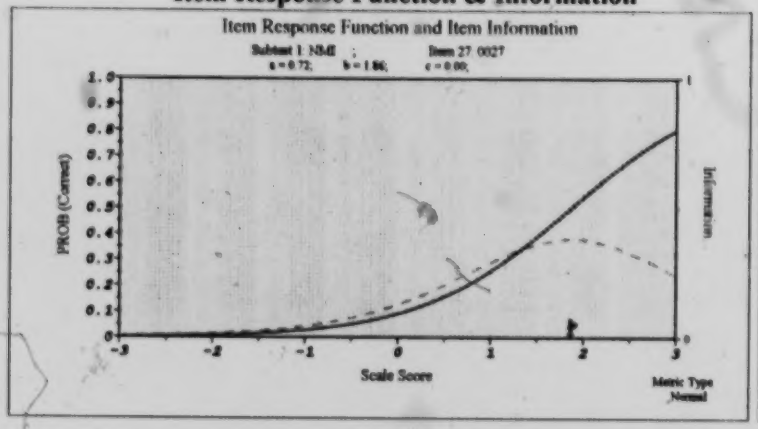


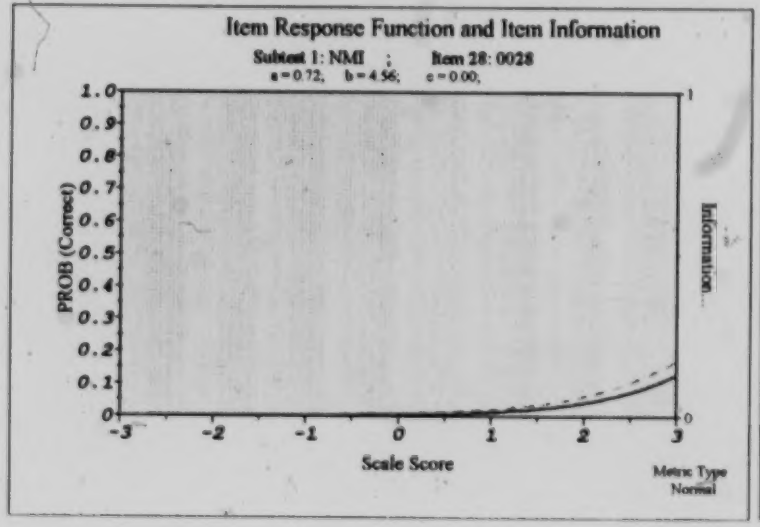
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 25: 0025
a = 0.72; b = 2.83; c = 0.00; chi-sq = 7.79; df = 5.00; prob < 0.167
PROB (Correct)
Scale Score
Metric Type
Normal

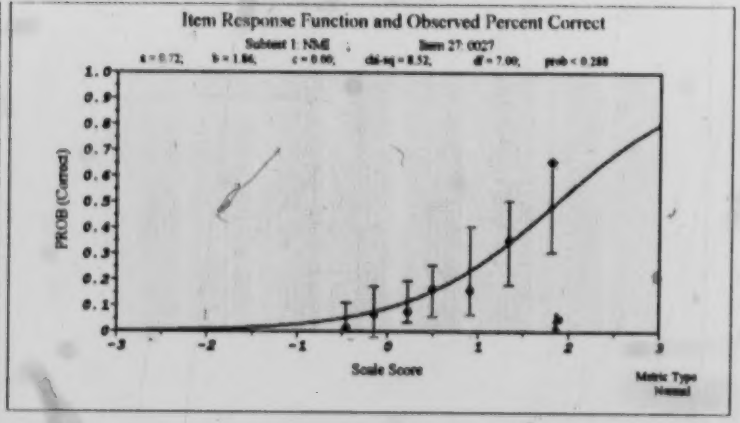

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 26: 0026
a = 0.72; b = 4.32; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

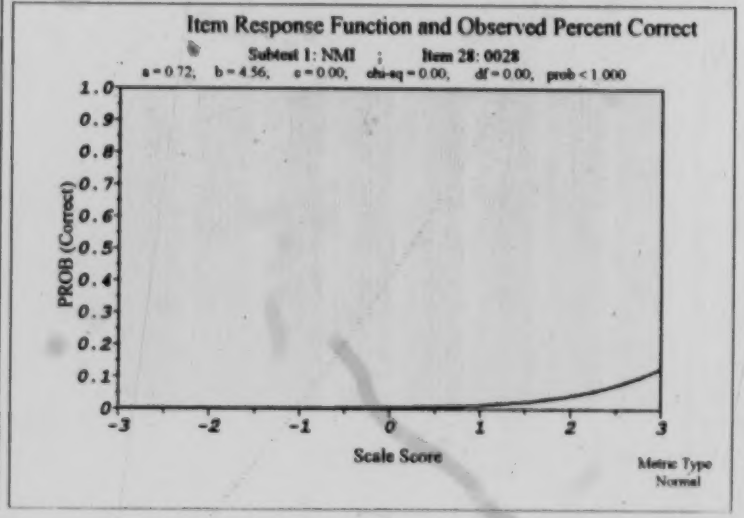
Item Response Function & Information



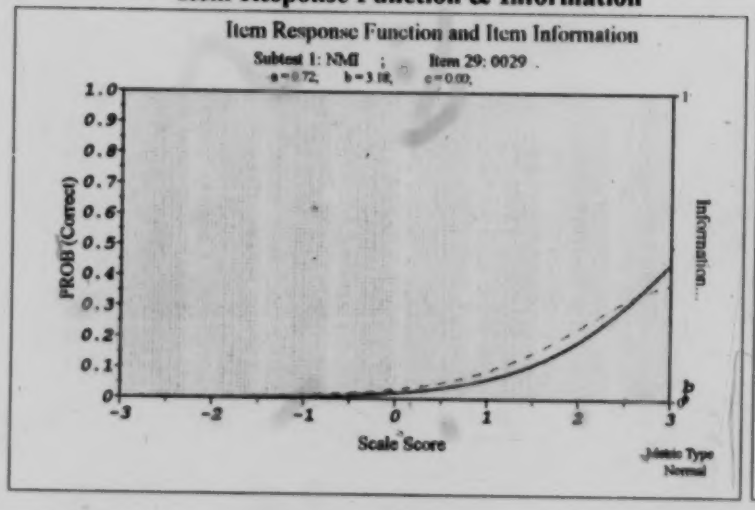


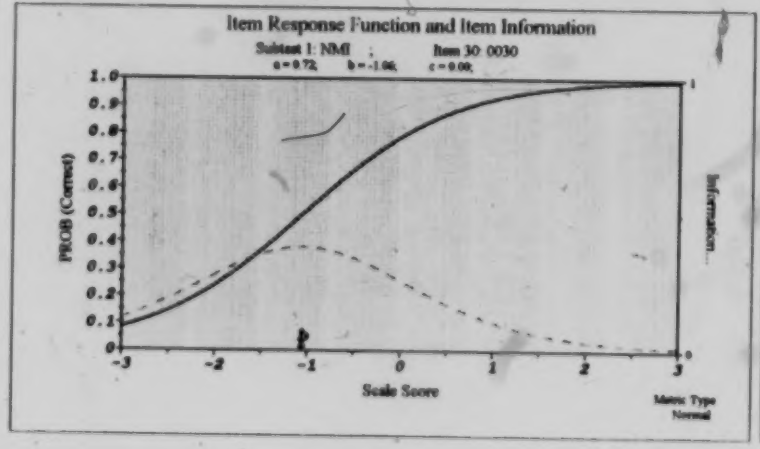
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 27: 0027
a = 0.72; b = 1.86; c = 0.00; chi-sq = 8.52; df = 7.00; prob < 0.288
PROB (Correct)
Scale Score
Metric Type
Normal

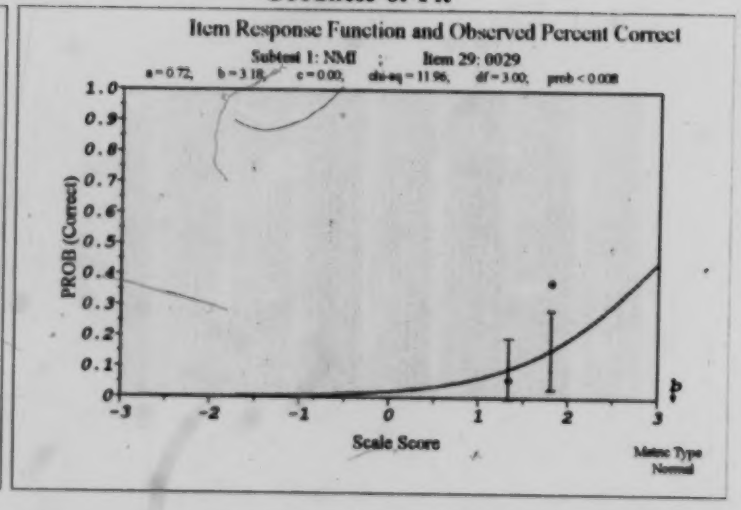

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 28: 0028
a = 0.72; b = 4.56; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

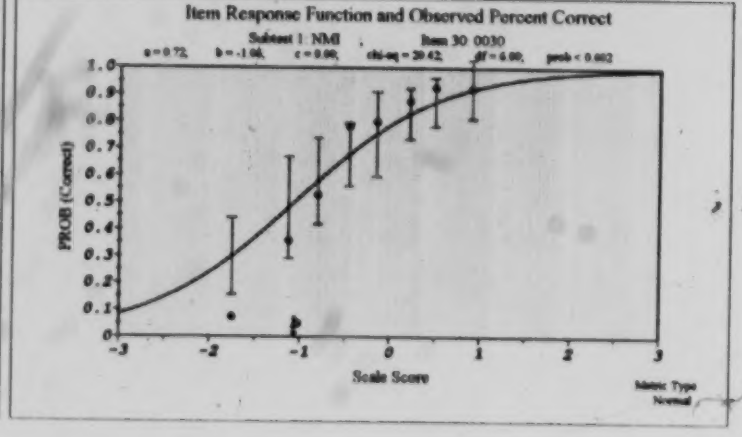
Item Response Function & Information



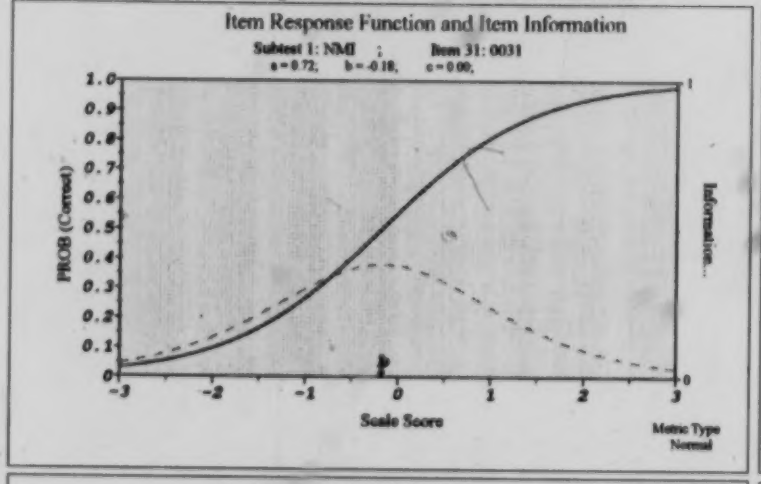


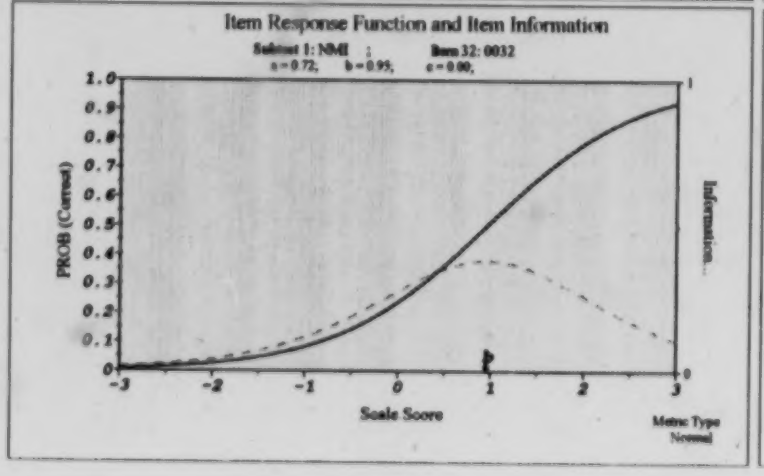
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 29: 0029
a = 0.72; b = 3.18; c = 0.00; chi-sq = 11.96; df = 3.00; prob < 0.008
PROB (Correct)
Scale Score
Metric Type
Normal

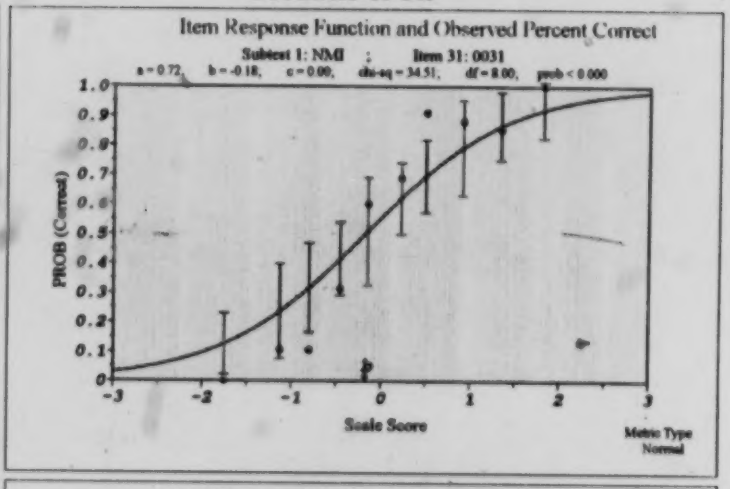

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 30: 0030
a = 0.72; b = -1.06; c = 0.00; chi-sq = 20.42; df = 6.00; prob < 0.002
PROB (Correct)
Scale Score
Metric Type
Normal

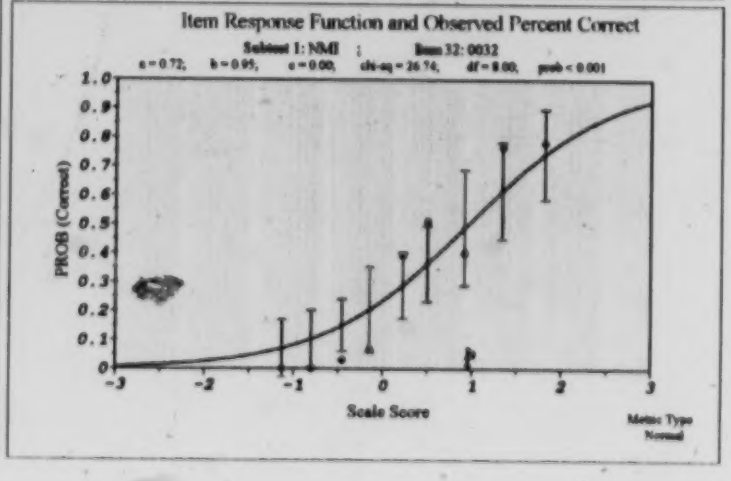
Item Response Function & Information



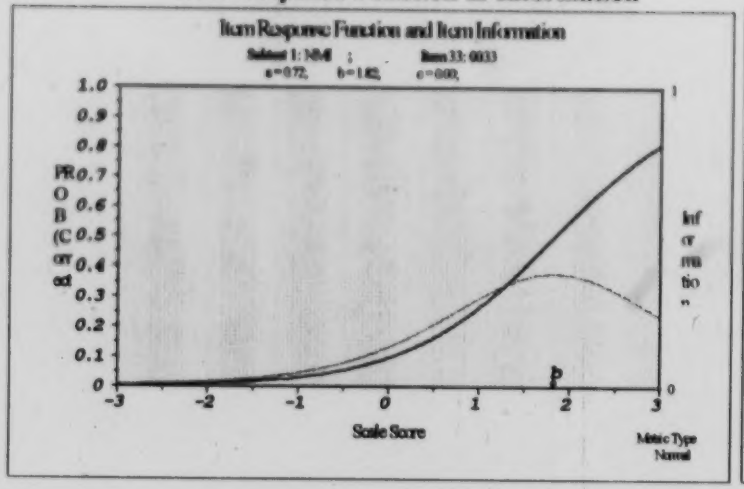


Goodness-of-Fit

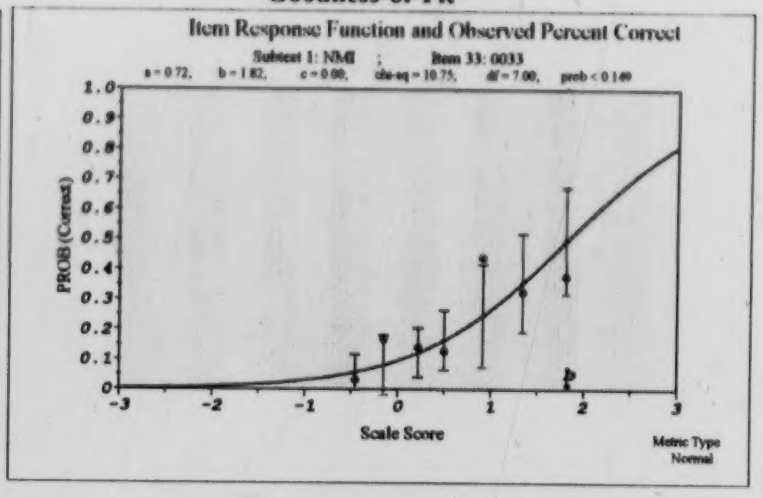

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 31: 0031
a = 0.72, b = -0.18, c = 0.00, chi-sq = 34.51, df = 8.00, prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

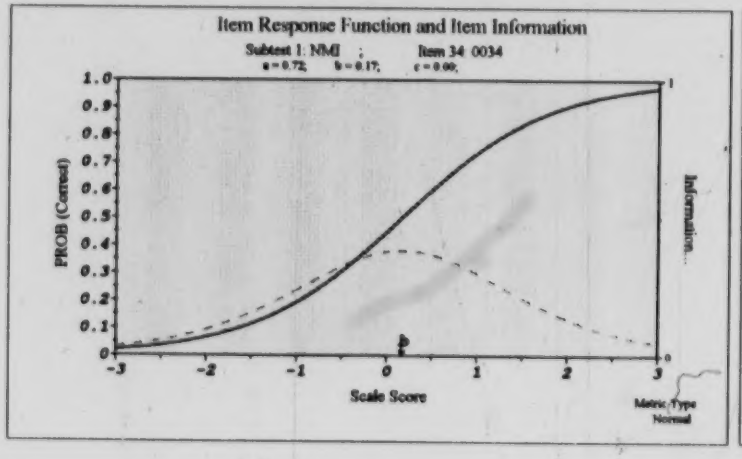

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 32: 0032
a = 0.72, b = 0.95, c = 0.00, chi-sq = 26.74, df = 8.00, prob < 0.001
PROB (Correct)
Scale Score
Metric Type
Normal

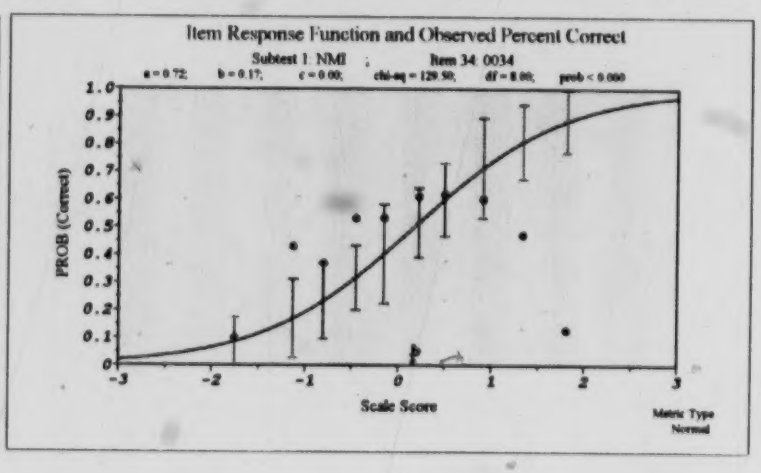
Item Response Function & Information



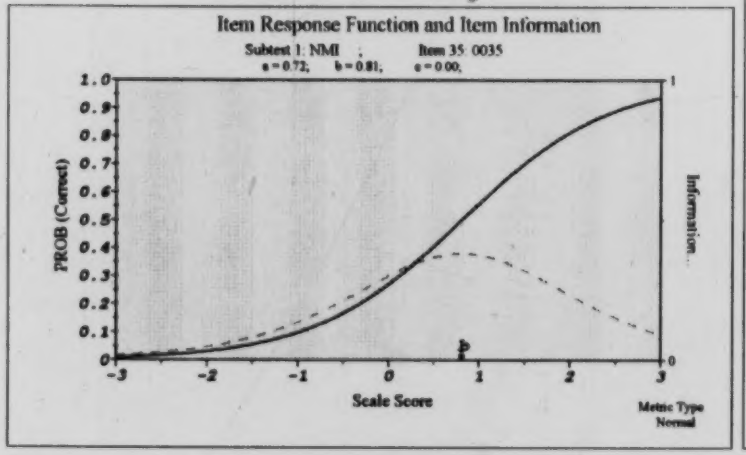
Goodness-of-Fit

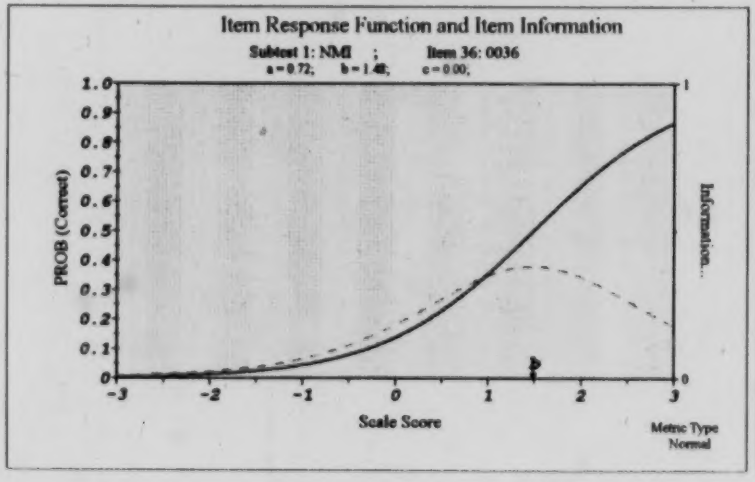

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 33: 0033
a = 0.72, b = 1.82, c = 0.00, chi-sq = 10.75, df = 7.00, prob < 0.149
PROB (Correct)
Scale Score
Metric Type
Normal



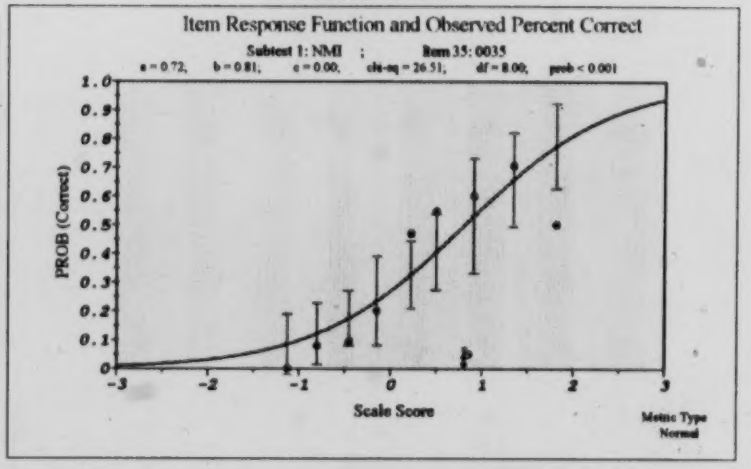

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 34: 0034
a = 0.72, b = 0.17, c = 0.00, chi-sq = 129.50, df = 8.00, prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

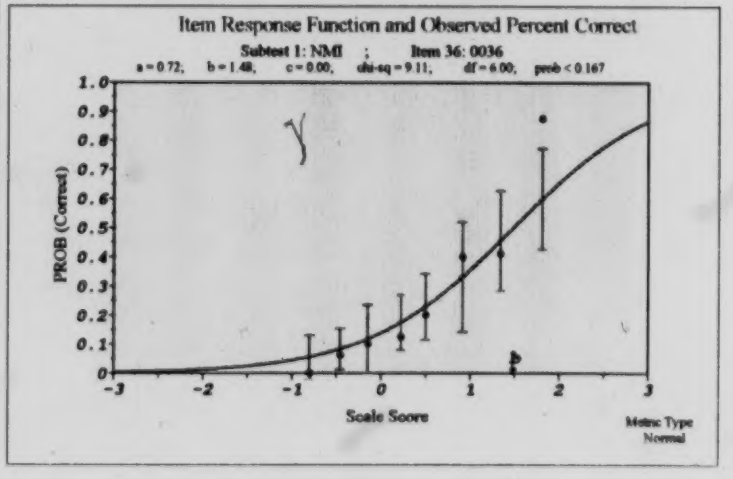
Item Response Function & Information



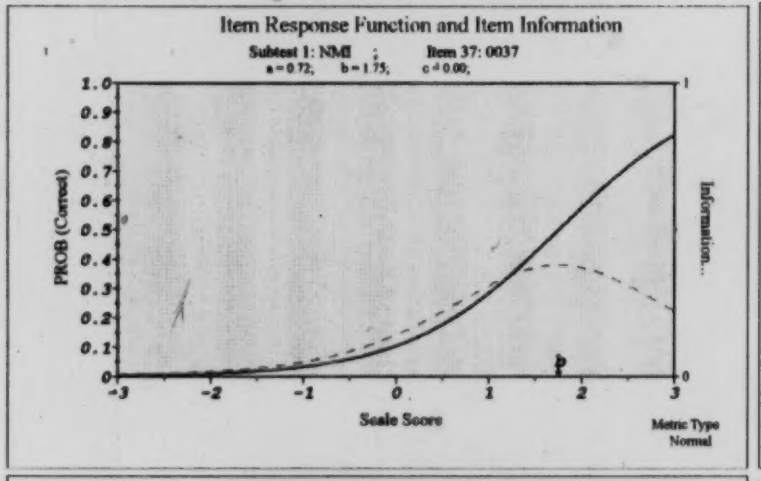


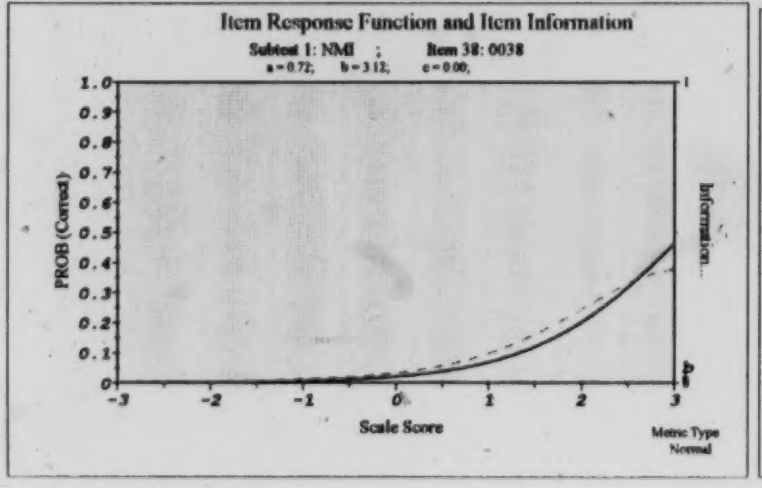
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 35: 0035
a = 0.72; b = 0.81; c = 0.00; chi-sq = 26.51; df = 8.00; prob < 0.001
PROB (Correct)
Scale Score
Metric Type
Normal

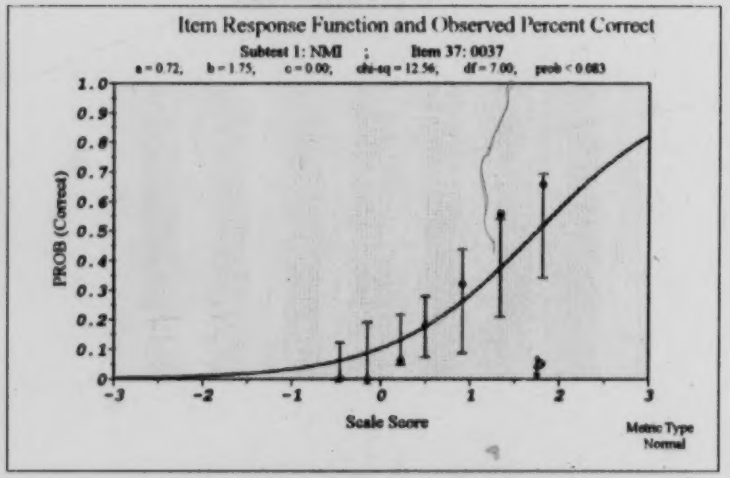

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 36: 0036
a = 0.72; b = 1.48; c = 0.00; chi-sq = 9.11; df = 6.00; prob < 0.167
PROB (Correct)
Scale Score
Metric Type
Normal

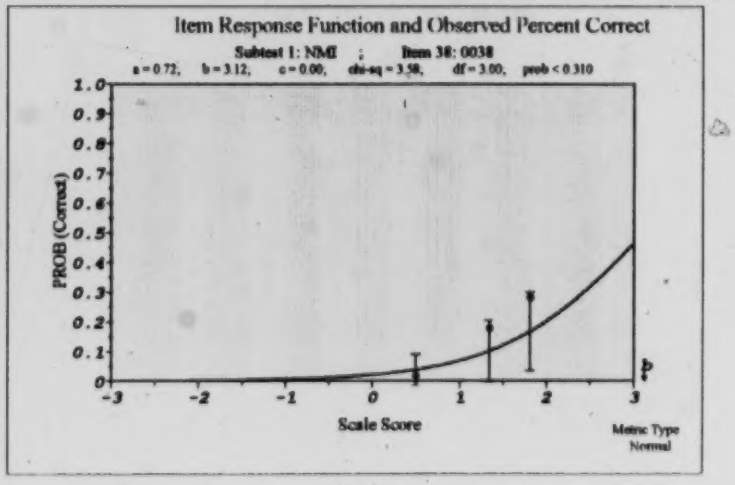
Item Response Function & Information



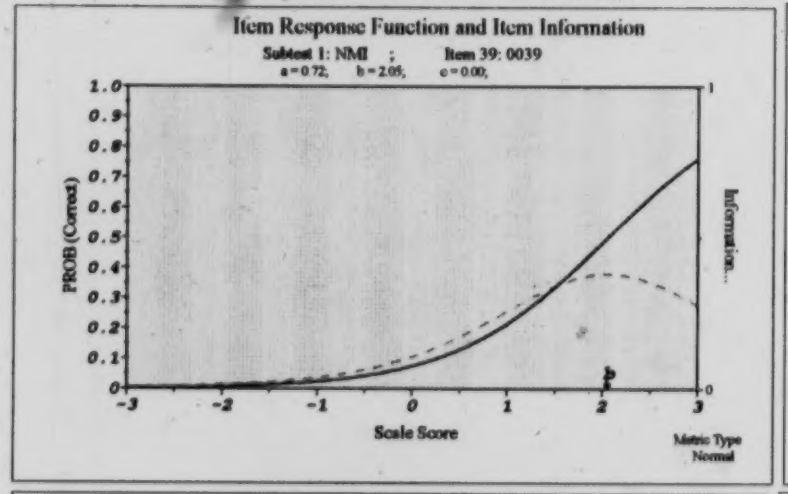


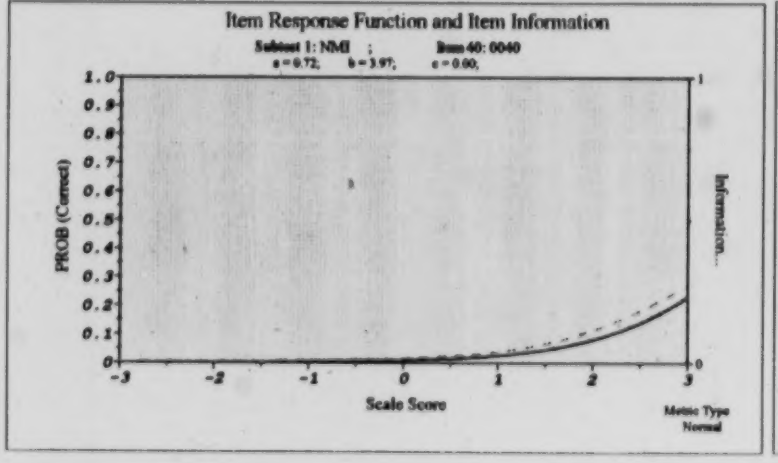
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 37: 0037
a = 0.72, b = 1.75, c = 0.00, chi-sq = 12.56, df = 7.00, prob < 0.083
PROB (Correct)
Scale Score
Metric Type
Normal

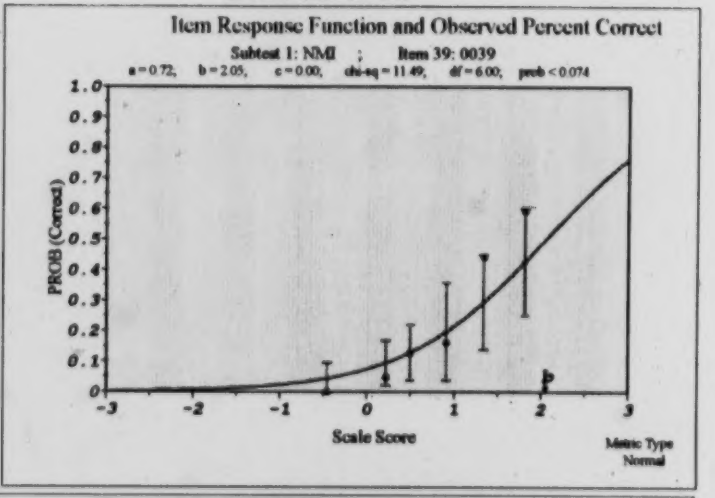

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 38: 0038
a = 0.72, b = 3.12, c = 0.00, chi-sq = 3.58, df = 3.00, prob < 0.310
PROB (Correct)
Scale Score
Metric Type
Normal

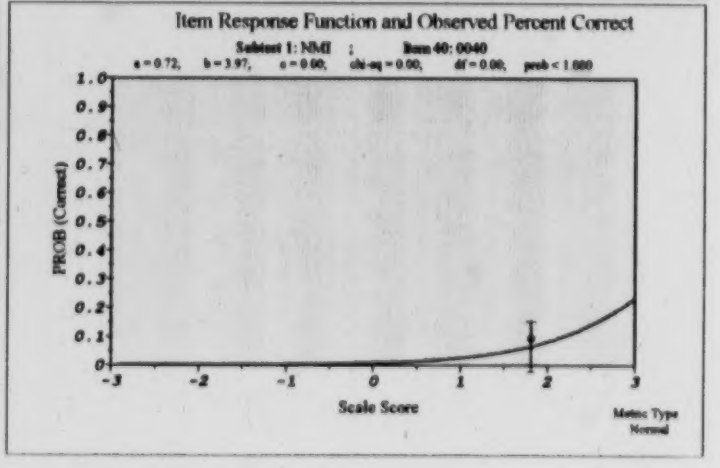
Item Response Function & Information



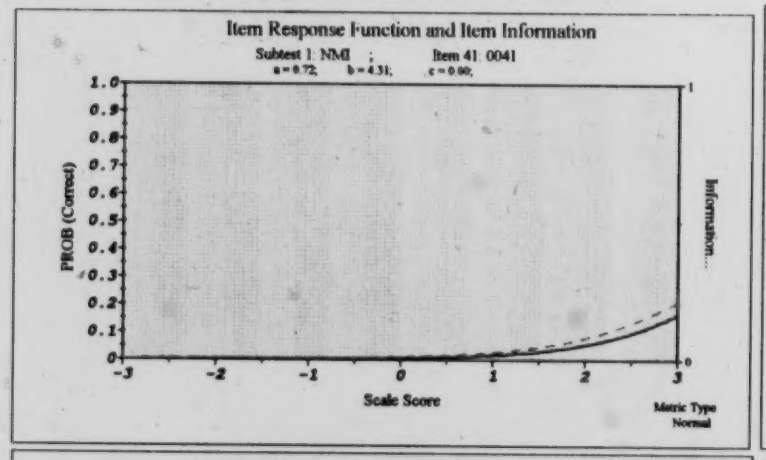


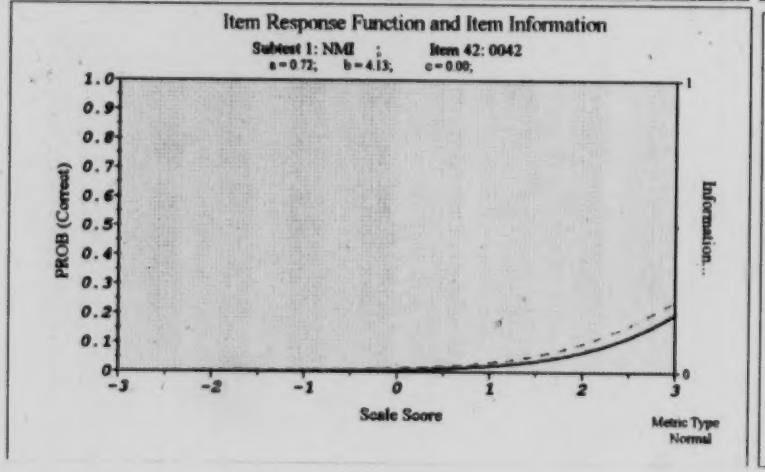
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 39: 0039
a = 0.72; b = 2.05; c = 0.00; chi-sq = 11.49; df = 6.00; prob < 0.074
PROB (Correct)
Scale Score
Metric Type
Normal

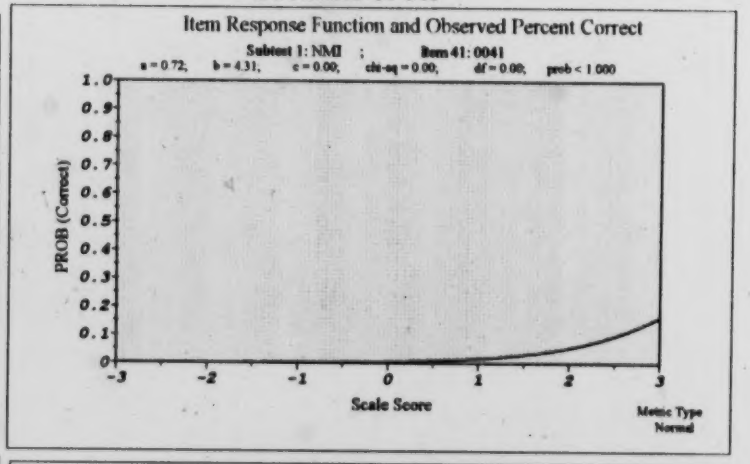

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 40: 0040
a = 0.72; b = 3.97; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

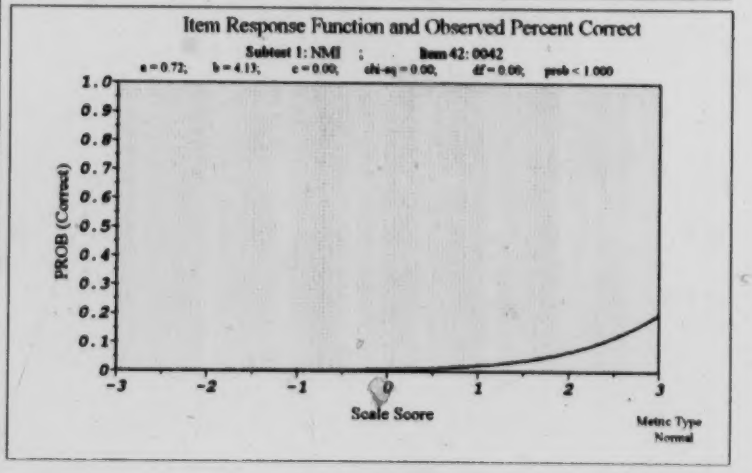
Item Response Function & Information



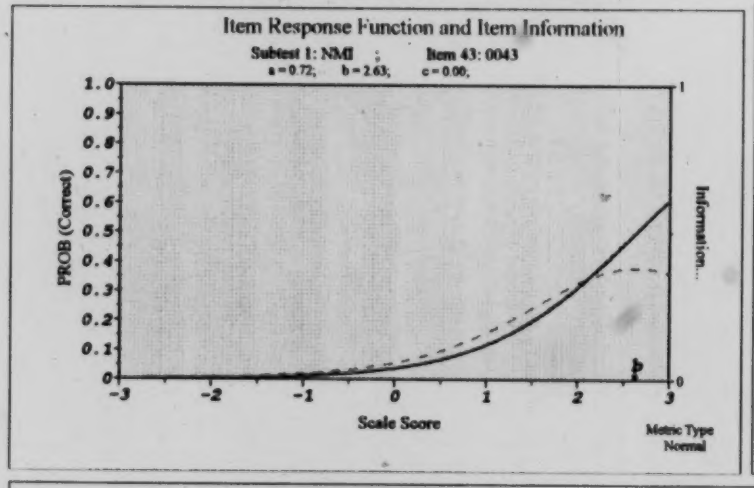


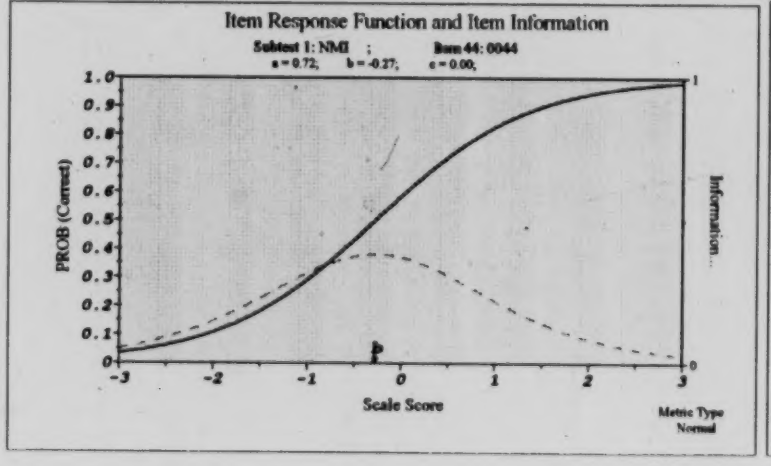
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 41: 0041
a = 0.72; b = 4.31; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

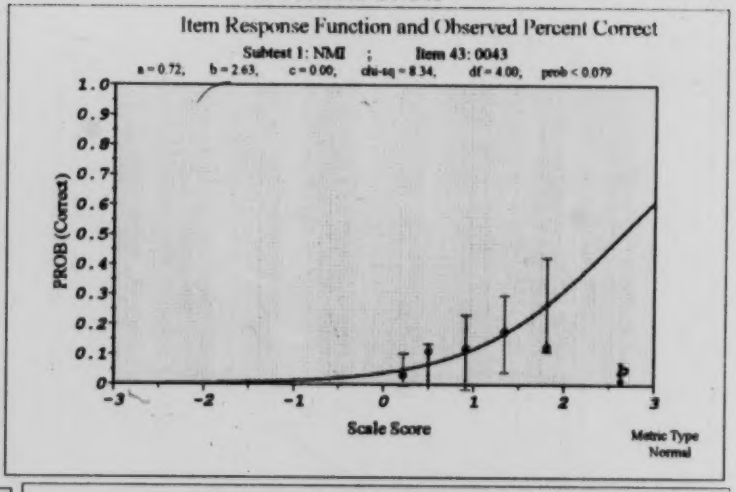

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 42: 0042
a = 0.72; b = 4.13; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

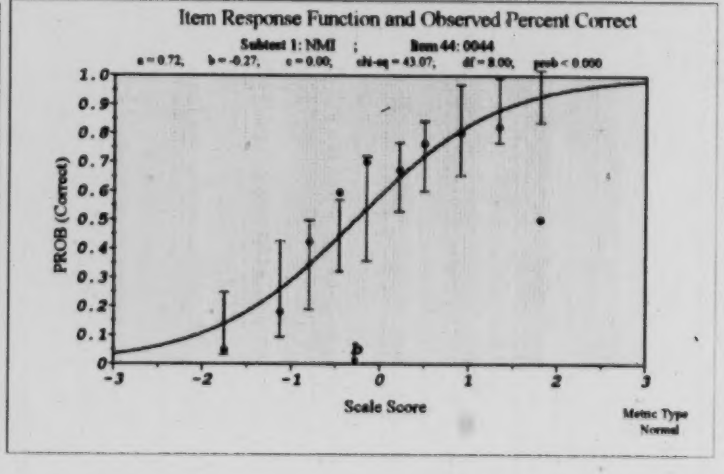
Item Response Function & Information



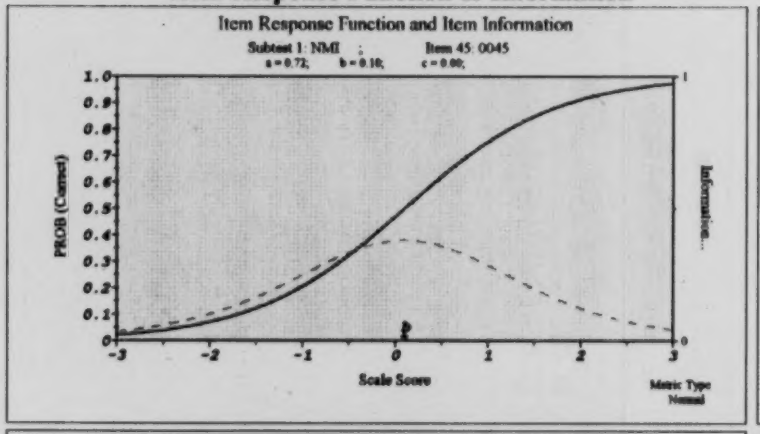


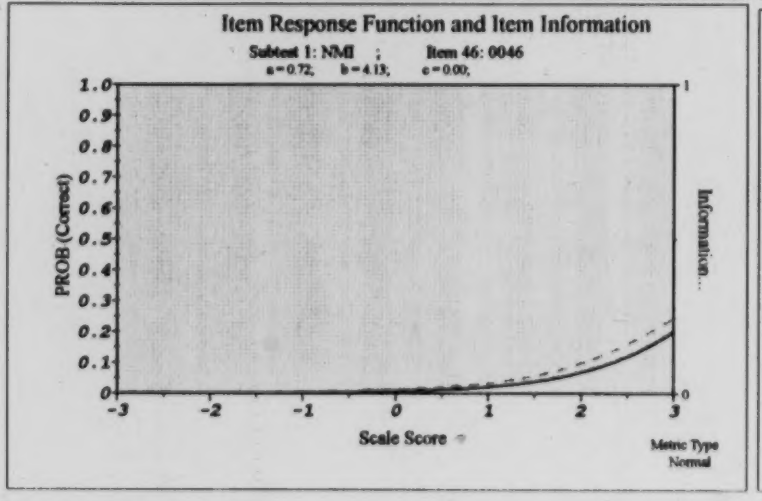
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 43: 0043
a = 0.72, b = 2.63, c = 0.00, chi-sq = 8.34, df = 4.00, prob < 0.079
PROB (Correct)
Scale Score
Metric Type
Normal

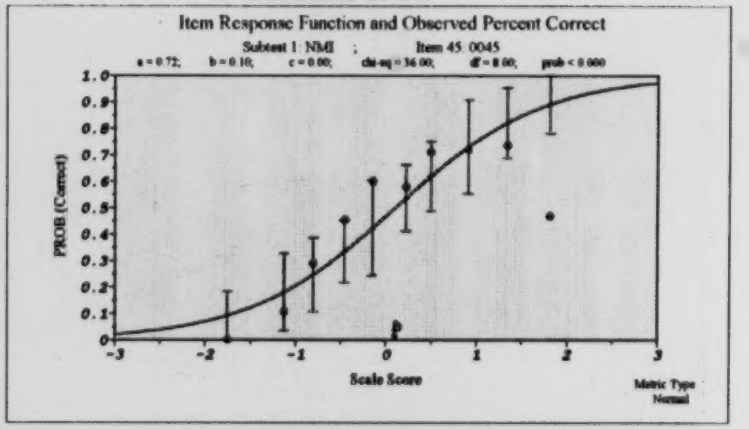

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 44: 0044
a = 0.72, b = -0.27, c = 0.00, chi-sq = 43.07, df = 8.00, prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

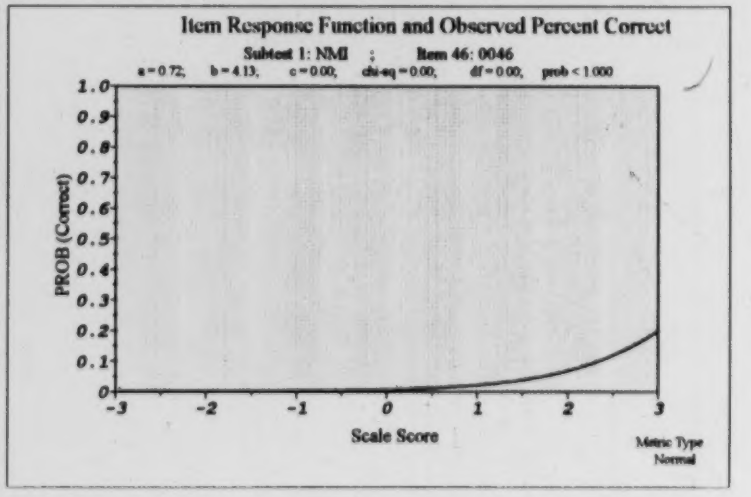
Item Response Function & Information



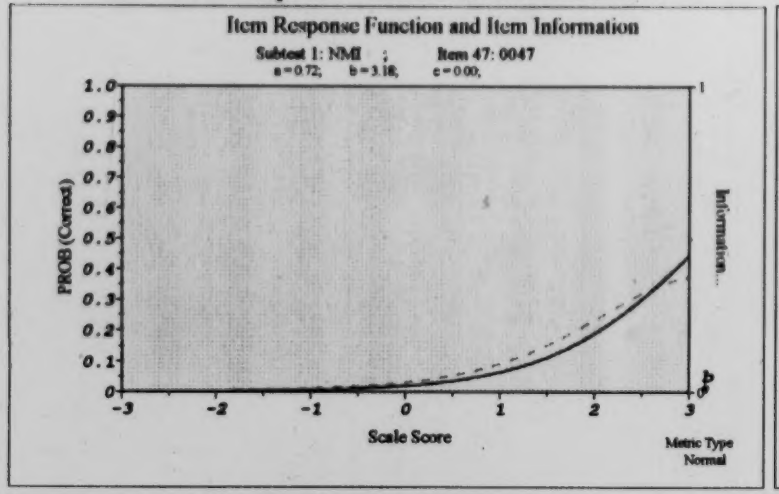


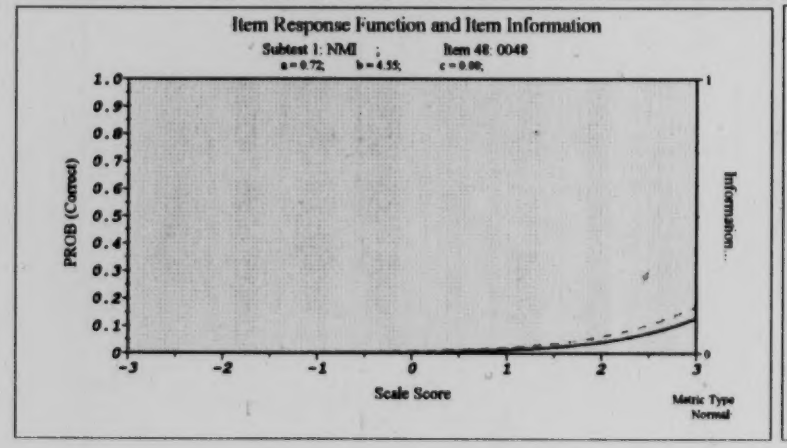
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 45: 0045
a = 0.72; b = 0.10; c = 0.00; chi-sq = 36.00; df = 8.00; prob < 0.000
PROB (Correct)
Scale Score
Metric Type
Normal

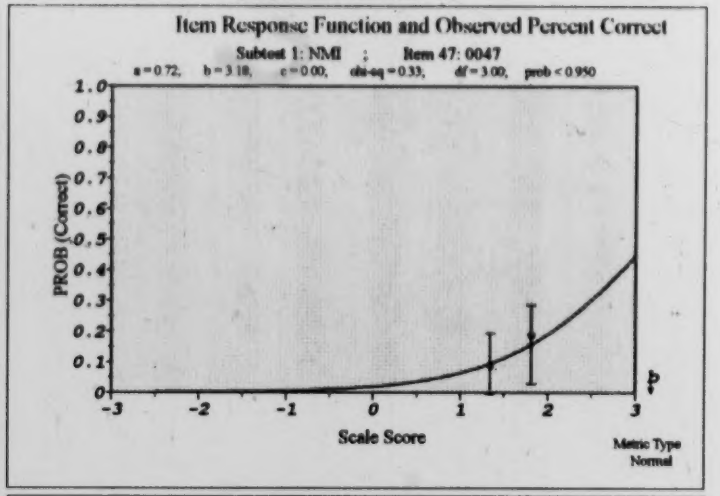

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 46: 0046
a = 0.72; b = 4.13; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

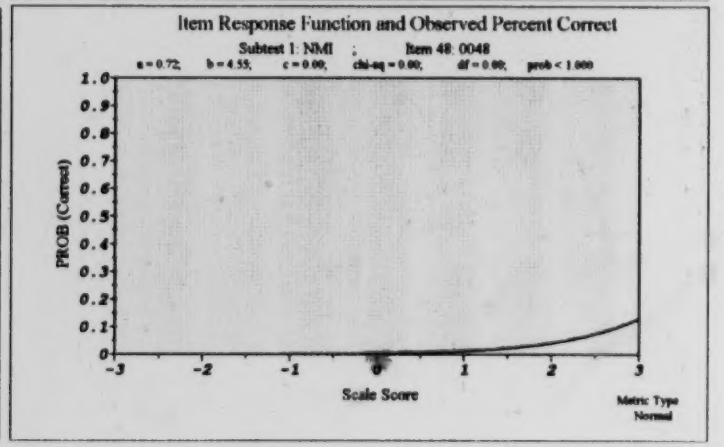
Item Response Function & Information



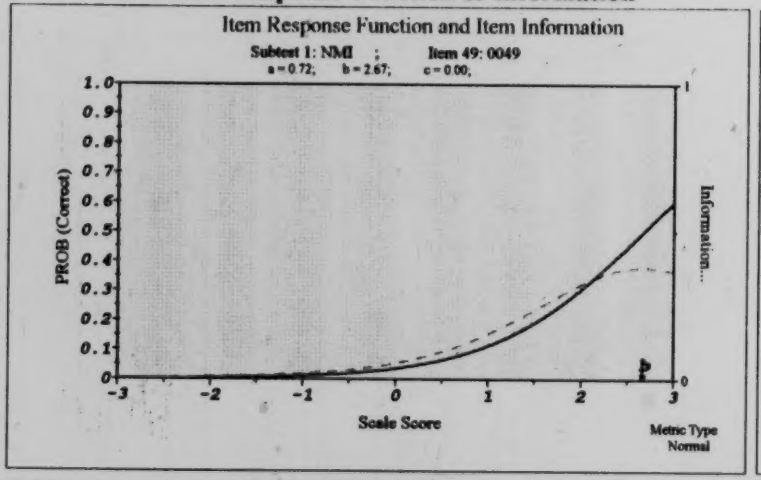


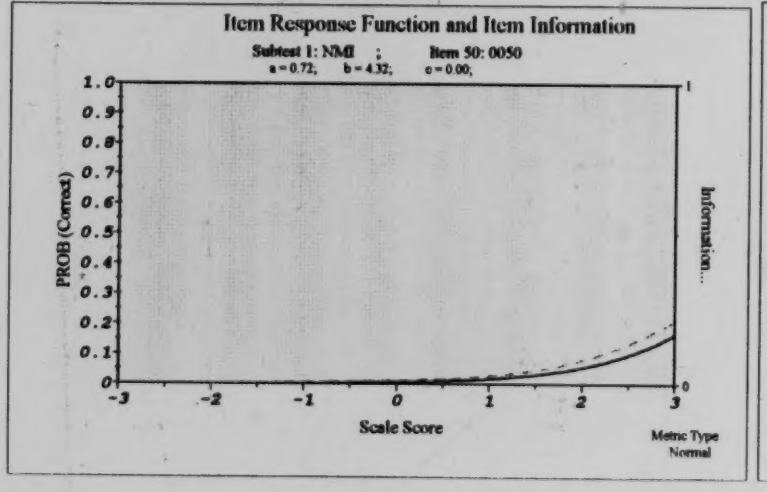
Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 47: 0047
a = 0.72; b = 3.18; c = 0.00; chi-sq = 0.33; df = 3.00; prob < 0.950
PROB (Correct)
Scale Score
Metric Type
Normal

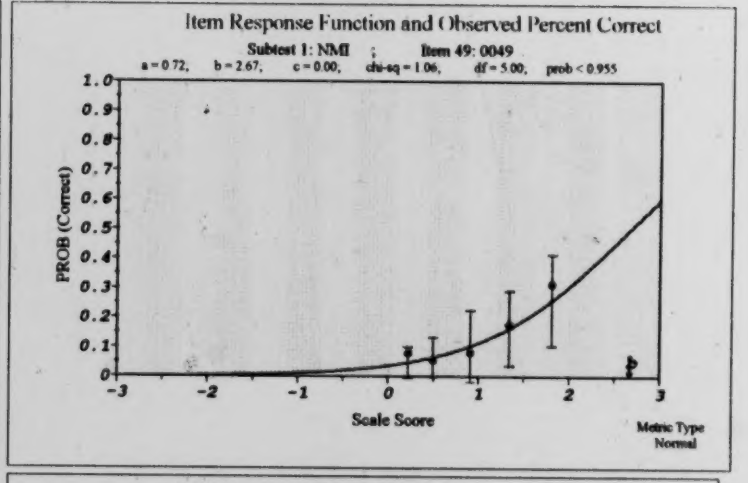

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 48: 0048
a = 0.72; b = 4.55; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

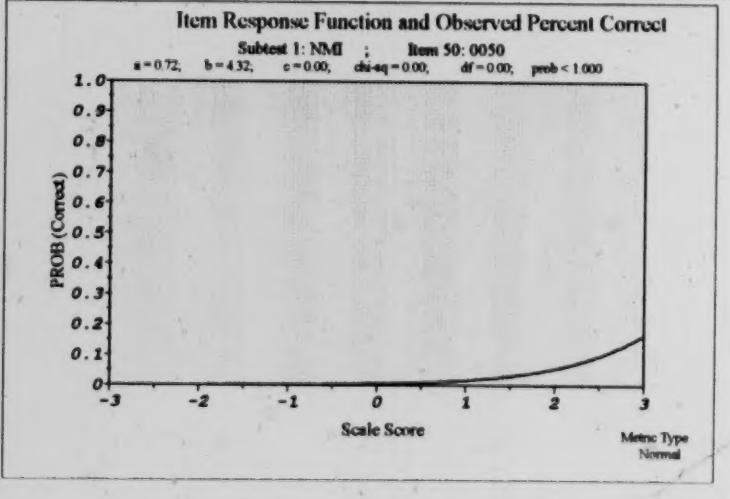
Item Response Function & Information





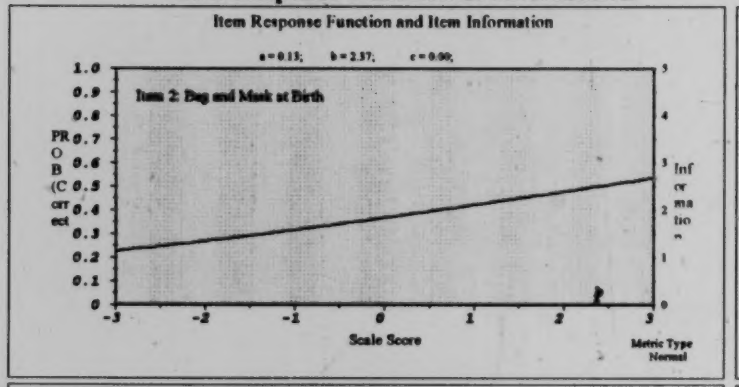
Goodness-of-Fit

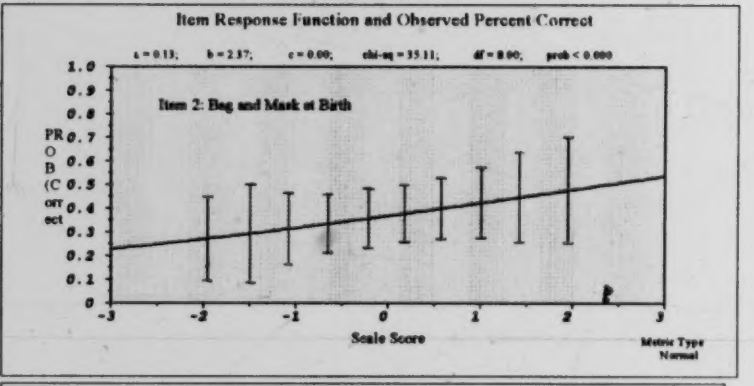

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 49: 0049
a = 0.72; b = 2.67; c = 0.00; chi-sq = 1.06; df = 5.00; prob < 0.955
PROB (Correct)
Scale Score
Metric Type
Normal

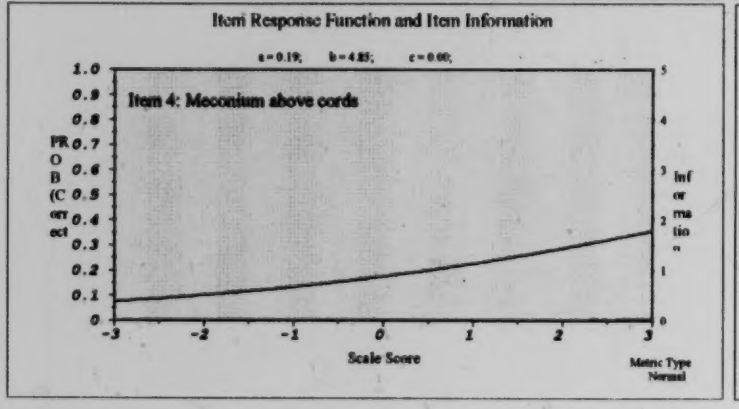

Item Response Function and Observed Percent Correct
Subtest 1: NMI ; Item 50: 0050
a = 0.72; b = 4.32; c = 0.00; chi-sq = 0.00; df = 0.00; prob < 1.000
PROB (Correct)
Scale Score
Metric Type
Normal

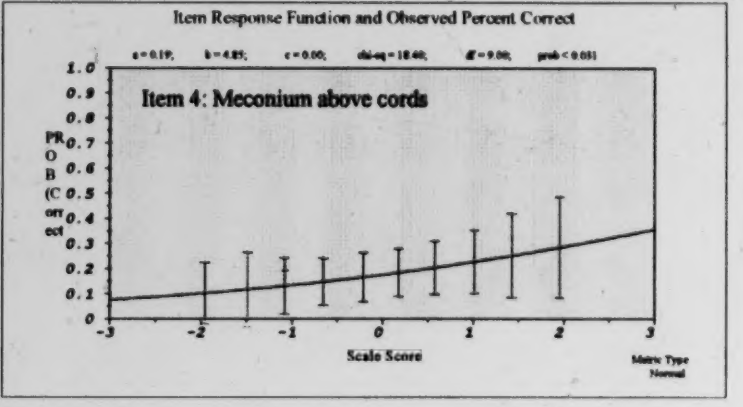
Appendix 11
Item Response Function Curves of Items with Poor-Fit [Deleted, n=7]
Refer Table 4-6, Column 58-Items BIN for Item Nos.

Item Response Function & Information | **Goodness-of-Fit**


Item Response Function and Item Information
a = 0.13; b = 2.57; c = 0.00;
Item 2: Bag and Mask at Birth
PROB (Correct)
Information
Scale Score
Metric Type Normal

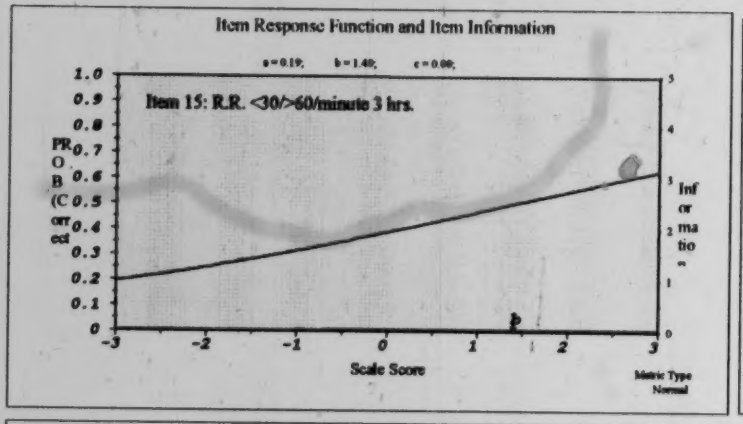

Item Response Function and Observed Percent Correct
a = 0.13; b = 2.57; c = 0.00; chi-sq = 35.11; df = 8.00; prob < 0.000
Item 2: Bag and Mask at Birth
PROB (Correct)
Scale Score
Metric Type Normal

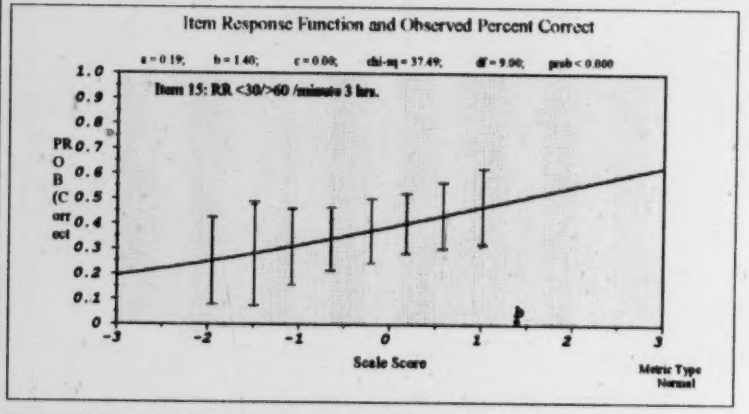

Item Response Function and Item Information
a = 0.19; b = 4.85; c = 0.00;
Item 4: Meconium above cords
PROB (Correct)
Information
Scale Score
Metric Type Normal

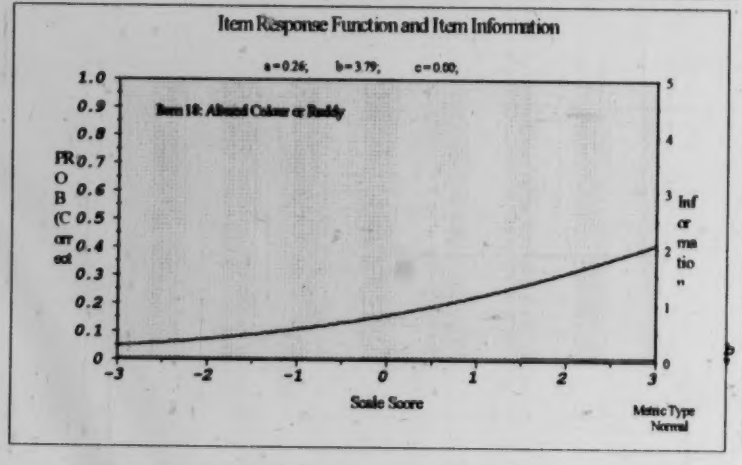

Item Response Function and Observed Percent Correct
a = 0.19; b = 4.85; c = 0.00; chi-sq = 18.40; df = 9.00; prob < 0.031
Item 4: Meconium above cords
PROB (Correct)
Scale Score
Metric Type Normal

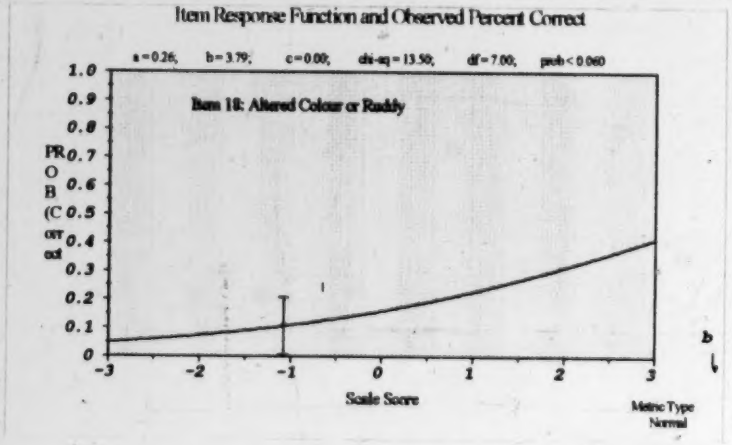
Item Response Function & Information

Goodness-of-Fit


Item Response Function and Item Information
a = 0.19; b = 1.40; c = 0.00;
Item 15: R.R. <30/>60/minute 3 hrs.
PROB (Correct)
Information
Scale Score
Metric Type Normal

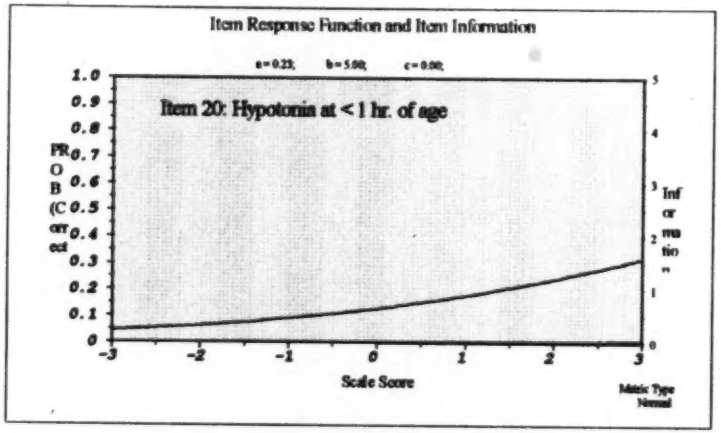

Item Response Function and Observed Percent Correct
a = 0.19; b = 1.40; c = 0.00; chi-sq = 37.49; df = 9.00; prob < 0.000
Item 15: RR <30/>60 /minute 3 hrs.
PROB (Correct)
Scale Score
Metric Type Normal

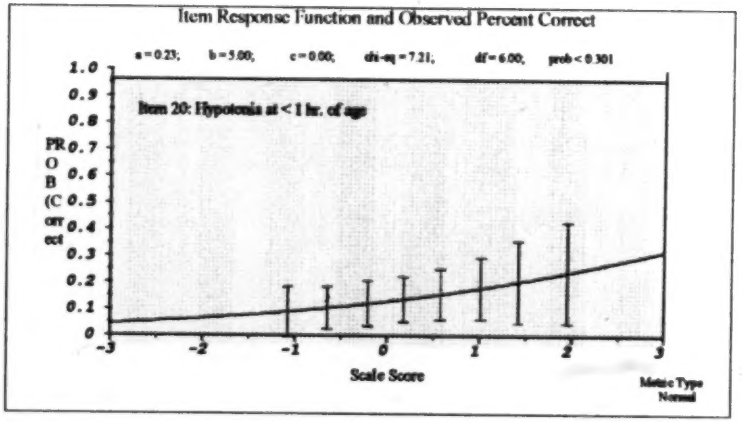

Item Response Function and Item Information
a = 0.26; b = 3.79; c = 0.00;
Item 18: Altered Colour or Ruddy
PROB (Correct)
Information
Scale Score
Metric Type Normal

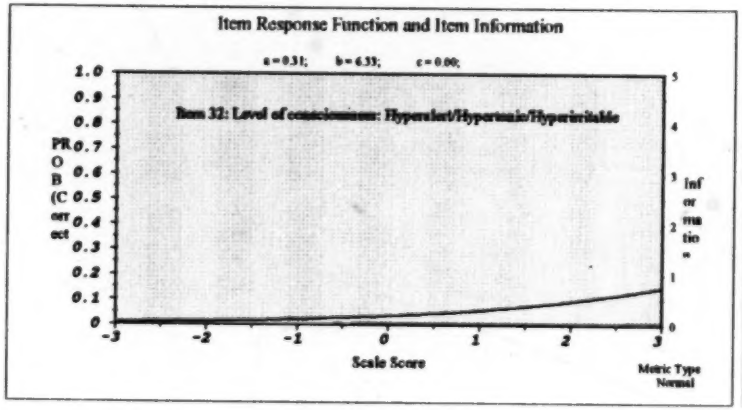

Item Response Function and Observed Percent Correct
a = 0.26; b = 3.79; c = 0.00; chi-sq = 13.50; df = 7.00; prob < 0.060
Item 18: Altered Colour or Ruddy
PROB (Correct)
Scale Score
Metric Type Normal

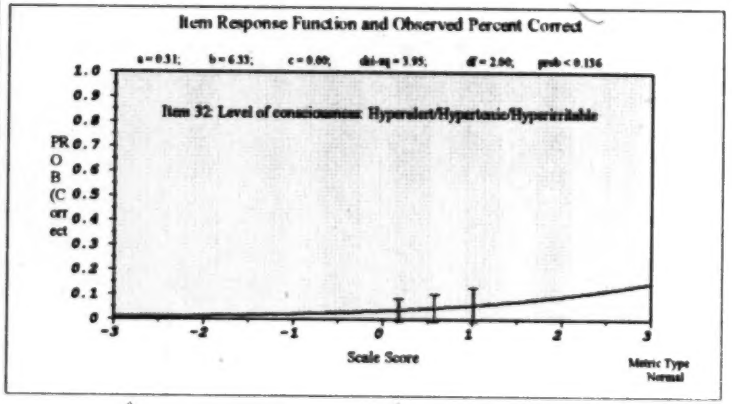
Item Response Function & Information

Goodness-of-Fit

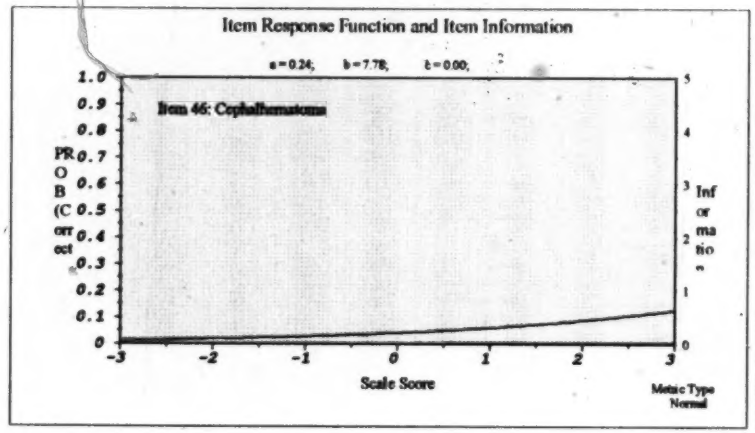

Item Response Function and Item Information
a = 0.23; b = 5.00; c = 0.00;
Item 20: Hypotonia at < 1 hr. of age
PROB (Correct)
Information
Scale Score
Metric Type Normal


Item Response Function and Observed Percent Correct
a = 0.23; b = 5.00; c = 0.00; chi-sq = 7.21; df = 6.00; prob < 0.301
Item 20: Hypotonia at < 1 hr. of age
PROB (Correct)
Scale Score
Metric Type Normal

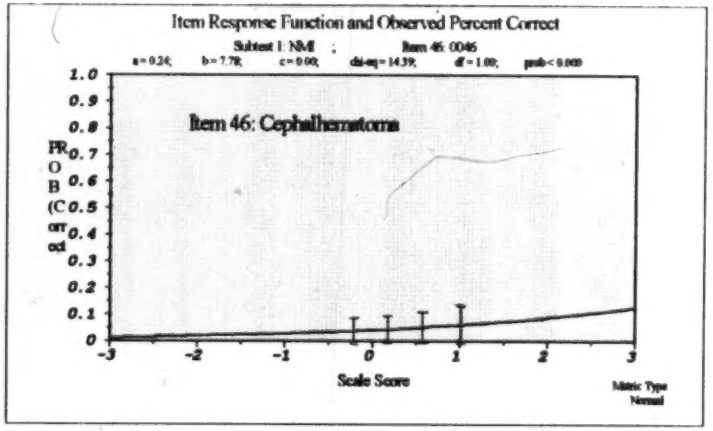

Item Response Function and Item Information
a = 0.31; b = 6.33; c = 0.00;
Item 32: Level of consciousness: Hyperalert/Hypertonic/Hyperirritable
PROB (Correct)
Information
Scale Score
Metric Type Normal


Item Response Function and Observed Percent Correct
a = 0.31; b = 6.33; c = 0.00; chi-sq = 3.95; df = 2.00; prob < 0.136
Item 32: Level of consciousness: Hyperalert/Hypertonic/Hyperirritable
PROB (Correct)
Scale Score
Metric Type Normal

Item Response Function & Information



Goodness-of-Fit


Item Response Function and Observed Percent Correct
Subtest I: NMI ; Item 46: 0046
a = 0.24; b = 7.78; c = 0.00; chi-sq = 14.39; df = 1.00; prob < 0.000
Item 46: Cephalhematoma
PROB (Correct
Scale Score
Metric Type Normal

Appendix 12 Study No. _____

Data Collection Form
50-Item Newborn Morbidity Index

1. Name
2. Gender
3. Gestational Age
4. Home Address and Phone
5. ID
6. Attending Physician
7. Date and Time of Birth

Date of Admission Date of Step-down Date of Discharge

Discharge Status

Dates Form Scored

AT BIRTH		WITHIN 24 HOURS (2 or more consecutive readings)	
Check	Morbidity Item	Check	Morbidity Item
	Cord Blood pH		Heart Rate/minute
1	<=7.1	9	161-200
	Resuscitation at Birth	10	>200
2	Intubation	11	<100
	Meconium		Systolic BP (mean, mm of Hg)
3	Meconium above and below cords		28-32 weeks 32-42 weeks
	Apgar Score (5 minute)	12	<30 <40
4	score<7		Respiratory Rate/minute
5	score<4	13	>100 between 3-24 hr.
6	score<1	14	<30 or >60 between 3-24 hr.
	Apgar Score (10 minute)		Altered Colour
7	score<7	15	Dusky OR Central Cyanosis
8	score<4		

Appendix 12 Study No. _____

Data Collection Form
50-Item Newborn Morbidity Index

BETWEEN BIRTH AND DISCHARGE FROM THE HOSPITAL			
Check	**Morbidity Item**	**Check**	**Morbidity Item**
	Hypotonia		**Respiratory Status**
16	Persisting (or identified) after 1 hr. of age	34	Assisted ventilation within 24 hr.
17	Persisting after 120 hr. of age	35	Assisted ventilation beyond 24 hr.
	Flaccidity	36	Mechanical ventilation within 24 hr.
18	Present (between 1-120 hr.)	37	Mechanical ventilation between 24 hr.- day 7
	Apnea	38	Mechanical ventilation beyond day 7
19	Apnea detected (by apnea monitor)		**Urine Output**
20	Apnea and need for oxygen	39	Low (< 2ml/ kg/ hour)
21	Apnea and need for resuscitation		**Birth Trauma**
	Bleeding Disorder	40	Fracture of Long Bone OR Clavicle OR Skull
22	Thrombocytopenia with or without	41	Nerve injury (Facial OR Peripheral)
	bleeding disorder (GI OR Lungs OR Skin)	42	Subdural OR Intracerebral hematoma
23	Need for transfusion due to Item 22		**Hypoglycemia (lowest level)**
	Seizures	43	Blood glucose <2.2 mmol/l
24	Single seizure		**Hyperbilirubinemia, micromol /L (peak level)**
25	>Single seizure (Multiple seizures)	44	Serum bilirubin >170
26	If >2 drugs used for treatment of seizures	45	Serum bilirubin >250 OR Phototherapy
	Level of Consciousness	46	Serum bilirubin >340 OR Exchange transfusion
27	Drowsy OR Lethargic		**Bacterial Culture**
28	Stupor OR Obtundation OR Coma	47	Blood positive
	Cardio-Pulmonary Resuscitation	48	CSF positive
29	Any time before discharge		**Intra-ventricular Hemorrhage**
	Oral Feeding Difficulties	49	Grade 1 OR 2
30	Poor sucking within 24hr.	50	Grade 3 OR 4
31	Poor sucking between 24 hr. - 7 days		
32	Poor sucking beyond day 7		
33	Persistent vomiting		